HABILITATION À DIRIGER DES RECHERCHES

SPÉCIALITÉ: MATHÉMATIQUES

Contribution à la statistique des diffusions. Estimation semiparamétrique et efficacité au second ordre. Agrégation et réduction de dimension pour le modèle de régression.

ARNAK S. DALALYAN

Rapporteurs :	M. Aad van der Vaart	vrije Universiteit (Amsterdam)
	M. Nakahiro Yoshida	University of Tokyo
	M. Oleg Lepski	Université de Provence

Soutenue le 22/11/2007 devant le jury composé de

M. Lucien Birgé	Université Paris 6
M. Jean Jacod	Université Paris 6
M. Youry Koutoyants	Université du Maine
M. Oleg Lepski	Université de Provence
Mme Dominique Picard	Université Paris 7
M. Alexandre Tsybakov	CREST

Contents

1

Overview

This document aims at summarizing my research activity in the theory of nonparametric and semiparametric statistics and shedding some light on the main ideas and tools leading to the results published in [P1]-[P10]. The three main directions of my research area are:

▷ statistical inference for continuously observed diffusion processes,

▷ second-order efficiency in semiparametric estimation,

▷ dimension reduction and aggregation in nonparametric regression with additive noise.

In this overview I will try to provide a brief account on the results obtained in each of these research areas. The purpose of this part is to discuss the results informally rather than stating rigorous mathematical assertions.

1.1 Inference for continuously observed diffusion processes

During my Ph.D. thesis, I started to work on the nonparametric inference for the model of continuously observed diffusion processes and, till now, it lies in the scope of my scientific interests. The general statistical problem can be formulated as follows. We have at our disposal one continuous curve $\boldsymbol{x}^T = \{x(t), 0 \leq t \leq T\}$ observed on the time interval $[0, T]$. This curve may be the time-continuous record of the stock price, the interest rate or some other random quantity varying continuously in time. We postulate that the curve we observed is a realization of a *time-homogeneous diffusion process* and we wish to make an inference on the parameters describing the stochastic dynamics of the underlying diffusion.

The dynamics of a time-homogeneous diffusion process is described by two functions: the instantaneous mean $S : \mathbb{R}_+ \to \mathbb{R}$ and the instantaneous variance $\sigma^2 : \mathbb{R}_+ \to \mathbb{R}$, which are re-

ferred to as *drift function* and *diffusion coefficient* respectively. Procedures for estimating these functions appeared in the statistical literature in the end of 1970ies [5, 40]. However, a more systematic study of nonparametric inference for diffusion processes has been initiated in the late 1990ies, stimulated by the impressive progress in the general theory of nonparametric statistics [27, 28, 31, 32, 47] and by the wide application of diffusion processes in finance [1, 3, 20, 22].

The theoretical criterion that has been mainly retained for assessing the quality of an estimating procedure was its aptitude to attain the asymptotically optimal/minimax rate of convergence when the time of observation T tends to infinity. In the same time, in the theory of nonparametric estimation for models having simpler stochastic structure, sharper results have been obtained pushing the theoretical study of optimality of statistical procedures up to the calculation of the optimal constants.

Thus, the goal of my Ph.D. thesis was to investigate whether it is possible or not to get asymptotic optimality/minimaxity up to the constant in the model of continuously observed diffusion processes. It turned out that the answer to this question is affirmative: we proved that the asymptotic minimaxity up to the optimal constant in the problem of estimating the drift function can be achieved by a proper choice of the kernel and the bandwidth of the kernel-type estimator. In the same time, we appropriately modified the L^2-risk serving as a measure of the quality of estimation as well as the functional class to which the unknown drift is supposed to belong to, by introducing a weight function equal to the square of the invariant density [P1, P2].

Two points should be stressed right away. First, since we assumed that a time-continuous record of a trajectory of a diffusion process is observed, the value of the diffusion coefficient at any point visited by x^T is computable using the quadratic variation. That is why we focused exclusively on the problem of estimating the drift function. Second, we investigated the case of positively-recurrent diffusion processes, considering thus only stationary processes. Note that this property guarantees the existence of the invariant density.

To construct an estimator of the drift, we used the fact that the value of the drift at some point x can be expressed as an algebraic function of the values at the point x of the diffusion coefficient, the invariant density and its derivative. Thus, we replaced the problem of estimating the drift by the problems of estimating the invariant density and its derivative. The treatment of these problems has been carried out in the same spirit as for estimating the density of iid observations, by virtu of the nice mixing properties of the underlying diffusion process.

Considering Sobolev-type smoothness classes, we obtained minimax results for estimating the derivative of the invariant density and the drift function. These results lie in the stream of the famous Pinsker theorem [41], which describes the asymptotic behavior, up to the constant, of the minimax risk in the Gaussian sequence model. However, instead of the global minimax approach of Pinsker, we adopted the local minimax approach which, in our opinion, leads to more elegant theoretical results in the problem of the drift estimation.

As a logical continuation, we addressed in [P4] the issue of the possibility of constructing an adaptive procedure attaining the asymptotically minimax bound of [P2]. In fact, the estima-

tor of the drift proposed in [P2] and proved to be asymptotically minimax up to the constant depends on the parameters of the functional class. Namely, if the drift is assumed to belong to a Sobolev ball, the computation of the estimator proposed in [P2] requires the knowledge of the smoothness index and the radius of the aforementioned Sobolev ball. Instead, the adaptive procedure constructed in [P4] does not depend neither on the smoothness index nor on the radius of the Sobolev ball and is asymptotically minimax simultaneously over a broad family of Sobolev balls. The construction of this adaptive procedure is essentially based on the method, going back to Mallows, of minimizing an unbiased risk estimate. The version that we used is inspired by the papers [24, 12].

In the context of invariant density estimation, a challenging issue was to investigate the second-order minimaxity of nonparametric estimators, in order to discriminate between different asymptotically first-order efficient estimators. In fact, it has been shown by Kutoyants [31] that the local time estimator, kernel-type estimators and a broad class of "unbiased estimators" are first-order asymptotically efficient. In [P3], we obtain a lower bound (up to the optimal constant) for the second-order minimax risk, and construct an estimator attaining this lower bound.

The results on the optimal constants and on the sharp adaptation for estimating the drift of a diffusion process appear to be very much in line with the analogous results in the classical nonparametric models (estimating a signal in Gaussian white noise, a regression function or a density of i.i.d. observations). This similarity advocates for a possible equivalence of the model of continuously observed diffusion with classical nonparametric models. Note that the long standing experience that under an asymptotic point of view the classical nonparametric models are statistically of the same kind has found its proper mathematical justification in 1996, when Brown and Low [8] and Nussbaum [39] proved the asymptotic equivalence of these models in the sense of Le Cam's theory of equivalent statistical experiments. In essence this means that any decision function developed for one model can be carried over, at least in an abstract way, to a decision function in the other models with exactly the same asymptotic risk properties. This is an important conceptual gain compared to the situation before where asymptotic results had to be proved each time separately.

In the papers [P5, P6], we showed strong asymptotic equivalence of the time-continuous diffusion model with a signal detection or Gaussian shift model, which can be interpreted as a regression model with random design. Our first results [P5] were established for the scalar diffusion model because we heavily employed tools from stochastic analysis that are neither available for time series analysis nor for multidimensional diffusion processes. More precisely, to prove the asymptotic equivalence we introduced a new coupling method providing an approximation of the likelihood of the diffusion model by a Gaussian one. The implementation of this idea was based on the local time of the diffusion process, which exists only in one-dimensional case.

Considering in a first step drift functions in a shrinking neighborhood of a known function S_0, we obtained local asymptotic equivalence results of the stationary diffusion experiment with, among others, an accompanying Gaussian regression experiment having the unknown drift as regression function and the invariant density associated to S_0 as design density. Note that the design can be considered random or deterministic in the sense that it determines the distance between two design points. This local asymptotic equivalence result has already

several implications for the statistical theory of diffusion processes. In particular, it can be used to obtain asymptotically sharp lower risk bounds. In order to transfer also global results like upper risk bounds to the diffusion case, a global equivalence result was obtained. In absence of a variance stabilizing transform the globally equivalent experiments are of compound type. Note that analogous results have been proved by Delattre and Hoffmann [19] for null recurrent diffusions having compactly supported drift. However, their arguments are well adapted to the null-recurrent case and seem to be inapplicable in the case of positively recurrent diffusions.

After the publication of our first paper [P5] on the asymptotic equivalence, we were frequently asked whether it is possible or not to adapt our coupling method to some settings where the local time does not exist. An answer to this question is given in [P6], where we show that the asymptotic equivalence between the diffusion model and the regression remains valid in the multi-dimensional case as well, in spite of the absence of local time. The main idea consists in including an additional space-discretization step allowing to replace the local time by the occupation measure.

Furthermore, similarly to [37], we established the asymptotic equivalence of the continuously observed diffusion model with the discretely observed diffusion with a step of discretization tending to zero at a suitable rate.

1.2 Second-order efficiency in semiparametrics

When I arrived at the University Paris 6, Sasha Tsybakov proposed me to join an ongoing project with Yuri Golubev having as target the study of second-order minimax properties of first-order efficient estimators in semiparametric statistics. To explain our motivation for tackling this problem, let me briefly recall some notions from semiparametric statistics.

In a semiparametric model, the parameter of interest is partitioned as $(\vartheta; f) \in \Theta \times \mathcal{F}$, with ϑ being a low-dimensional parameter of interest and f a higher dimensional (often infinite-dimensional) nuisance parameter. A popular method of estimating ϑ for unknown f is the profile likelihood maximization [46, 52]. Let $l_n(\vartheta; f)$ be the log-likelihood of the model, the profile likelihood for ϑ is defined as $pl_n(\vartheta) = \sup_{f \in \mathcal{F}} l_n(\vartheta; f)$ and the Profile Likelihood Estimator (PLE) is $\vartheta_{PLE} = \arg\max_\vartheta pl_n(\vartheta)$. Thus, the nuisance parameter f is eliminated by taking the supremum over all possible values of f in some a priori chosen class $\mathcal{F}$ assumed to contain the true value of f. Using tools from the empirical process theory, Murphy and van der Vaart [38] proved that, under mild assumptions, the PLE is semiparametrically first-order efficient.

Since often $\mathcal{F}$ is infinite-dimensional, the maximization in $f \in \mathcal{F}$ may be difficult to perform both from theoretical and practical points of view. A useful idea is therefore to regularize this optimization problem either by replacing $\mathcal{F}$ by a finite-dimensional set or by penalizing the likelihood, or by using another smoothing technique.

A natural question arises: what is the best regularization and what is its impact on the accuracy of the resulting estimator? The theory fails to give a complete answer to this question

as long as only the first-order term of the risk is considered. Usually, and it is also the case for the model of shift estimation for a periodic signal corrupted by a Gaussian white noise [P7, P8], there is a large variety of regularization methods leading to first-order efficient estimators. A particularly appealing way to choose the "best" estimator among these first-order efficient estimators consists in comparing the second-order terms of their "worst case" risks. This leads to the second-order minimax approach which has been firstly developed by Golubev and Härdle [25, 26] for partial linear models. The techniques used in [25, 26] hardly rely on the linearity of the model on the parameter of interest ϑ. We were therefore interested in extending this approach to models having nonlinear structure.

Thus, in [P7, P8], we developed the second-order minimax approach for the model of shift estimation of a periodic signal corrupted by Gaussian white noise. This is an "idealization" of the symmetric location problem, which is often considered as a prototype in semiparametric inference [48, 49]. As we stressed in [P7], the aforementioned model seems to capture main difficulties in deriving second-order efficiency, being at the same time simple enough to avoid irrelevant technicalities. A partial confirmation of this conjecture are the results of Castillo [11] who, following the general scheme described in [P7], proved quite similar results for the problem of estimating the scaling parameter of a signal corrupted by Gaussian white noise. While in [P7] second-order efficient estimators of the shift parameter were proposed in the case where the signal belongs to a Sobolev ball with known smoothness index and radius, the aim of [P8] was to construct a second-order efficient estimator which is entirely data-dependent. We achieved this aim by using a penalized profile likelihood estimator based on Stein's blockwise shrinkage idea and we proved its second-order minimaxity simultaneously for a large scale of Sobolev balls.

The results of the above mentioned papers grant an increasing importance to the second-order efficiency in that they show that, in a semiparametric estimation problem, the second-order term is not dramatically smaller than the first-order term, especially when the nuisance parameter is not very smooth. Thus, finding second-order efficient estimators is not only a challenging theoretical problem, but is also of some practical interest.

1.3 Dimension reduction and aggregation in nonparametric regression

The regression with additive noise is certainly one of the most studied models in statistics. In spite of this, there are still many challenging open problems related to this model especially in the case where the explanatory variable is high-dimensional. Usually, statistical procedures designed to work for a relatively large nonparametric class of regression functions exhibit poor empirical performance. To elaborate more performant statistical procedures some additional structural assumptions are to be imposed.

The problem we are concerned with is to predict or to explain a response variable Y by d scalar covariates $X^{(1)}, \dots, X^{(d)}$. To accomplish this task, the only thing we have at our disposal is a sample of size n of these variables. Assume for the moment that there is a function f characterizing the relationship between Y and $\boldsymbol{X} = (X^{(1)}, \dots, X^{(d)})^\top$. The papers

[P9, P10] deal with the regression model with additive noise under different assumptions on the form of f. The main aim is to design statistical procedures that take advantage of the specific assumption and to infer theoretical results on their performance.

1.3.1 Dimension reduction in multi-index model

We consider the multi-index model with m^* indices: for some linearly independent vectors $\vartheta_1, \ldots, \vartheta_{m^*}$ and for some function $g : \mathbb{R}^{m^*} \to \mathbb{R}$, the relation $f(\boldsymbol{x}) = g(\vartheta_1^\top \boldsymbol{x}, \ldots, \vartheta_{m^*}^\top \boldsymbol{x})$ holds for every $\boldsymbol{x} \in \mathbb{R}^d$. Here and in the sequel the vectors are understood as one column matrices and $M^\top$ denotes the transpose of the matrix M. Of course, such a restriction may lead to a substantial improvement of statistical performance of procedures only if $m^* < d$. The main argument in favor of using the multi-index model is that for many data sets the underlying structural dimension m^* is much smaller than d. Therefore, if the vectors $\vartheta_1, \ldots, \vartheta_{m^*}$ are known, the estimation of f reduces to the estimation of g, which can be performed much better because of lower dimensionality of the function g compared to that of f.

Another advantage of the multi-index model is that it assesses that only few linear combinations of the predictors may suffice for "explaining" the response Y. Considering these combinations as new predictors leads to a much simpler model (due to its low dimensionality), which can be successfully analyzed by graphical methods, see [18, 15] for more details.

Since it is unrealistic to assume that $\vartheta_1, \ldots, \vartheta_{m^*}$ are known, the estimation of these vectors from the data is of high practical interest. When the function g is unspecified, only the linear subspace $\mathcal{S}_\vartheta$ spanned by these vectors may be identified from the sample. This subspace is usually called *index space* or *dimension-reduction (DR) subspace*. Clearly, there are many DR subspaces for a fixed model f. Even if f is observed without error, only the smallest DR subspace, henceforth denoted by $\mathcal{S}$, can be consistently identified. This smallest DR subspace, which is the intersection of all DR subspaces, is called *effective dimension-reduction* (EDR) subspace [35] or *central mean subspace* [16].

When I was postdoc in Berlin, Volodia Spokoiny introduced me to the method of structural adaptation and its applications in multivariate statistics. In particular, in [29], this method has been used to estimate the EDR subspace. The main idea was to exploit the fact that the gradient ∇f of the regression function f evaluated at any point $\boldsymbol{x} \in \mathbb{R}^d$ belongs to the EDR subspace in order to construct some vectors $\beta_1, \ldots, \beta_L$ nearly lying in the EDR subspace, and to estimate a basis of the EDR subspace by means of the Principal Component Analysis (PCA).

A limitation of this method is that the resulting estimator is proved to be $\sqrt{n}$-consistent only when L is chosen independently on the sample size n. Unfortunately, if L is small with respect to n, it is hopeless that the subspace spanned by the vectors $\beta_1, \ldots, \beta_L$ captures all the directions of the EDR subspace. Therefore, the empirical experience advocates for large values of L, even if the desirable feature of $\sqrt{n}$-consistency fails in this case.

The goal of [P9] was to propose an estimator providing a remedy for this dissension between the theory and the empirical experience. To this end, we introduced a new method of extracting the EDR subspace from the vectors $\beta_1, \ldots, \beta_L$. If we think of PCA as the solution

to a minimization problem involving a sum over L terms then, to some extent, our proposal was to replace the sum by the maximum. This is why we called our procedure Structural Adaptation via Maximum Minimization (SAMM).

The main advantage of SAMM was that it is proved to give a consistent estimator of the EDR subspace under a very weak identifiability assumption, even in the case where L is of polynomial order in n. In addition, the rate of convergence of the proposed estimator is $\sqrt{n}$ (up to a logarithmic factor) when $m^* \leq 4$. We also studied the numerical performance of SAMM by means of Monte Carlo simulations. The results presented in [P9, Section 4] show the state-of-the-art performance of SAMM.

1.3.2 Aggregation and sparsity oracle inequalities

The method proposed in [P9] provides an estimator of the EDR subspace in the case where its dimension is known. This allows one to use standard nonparametric smoothing techniques in order to define estimators of the regression function, after projecting the covariates onto the estimated EDR subspace. Since the estimation of the EDR subspace may be done with parametric rate of convergence, the resulting nonparametric estimators of the regression function will have the same rate as those using the projection onto the true EDR subspace. This rate will depend on the underlying structural dimension (dimension of the EDR subspace) and not on the real dimension of the explanatory variable. An important limitation here is that this construction presumes the knowledge of the structural dimension.

One possible approach for overcoming this difficulty passes through the aggregation of regression estimators. In our work, we only consider convex aggregation, the purpose of which can be formulated as follows: having at hand the data $\mathcal{D}$ and a collection of estimators $\mathcal{F}$, choose an element (called aggregate) in the convex hull of $\mathcal{F}$ which is nearly as close to the true regression function as the best estimator from $\mathcal{F}$. Thus, a possible strategy for efficiently estimating the regression function in the multi-index model without knowing the structural dimension consists in building in a first step regression estimators for every possible value of the structural dimension and, in a second step, aggregating these estimators to obtain an estimator of the regression function which is adaptive with respect to the unknown structural dimension.

To realize this program, we were looking for results on aggregation of estimators in the model of regression with deterministic design. Surprisingly, most of results on aggregation were concerned with the model of regression with random design. Yang [54, Remark 4 on page 151] even questions whether the results on aggregation for random design regression may be carried over the regression with deterministic design. To the best of our knowledge, the only paper where this issue is addressed is that of Leung and Barron [34]. It should be stressed here that the results of [34] are particularly remarkable given that they provide sharp oracle inequalities for the aggregate with exponential weights defined without sample-splitting.

However, a limitation of Leung and Barron's results is that they are heavily based on the assumption that the regression errors are normally distributed. In the same time, Juditsky,

Rigollet and Tsybakov [30] give an elegant proof of a sharp oracle inequality (quite similar to that of [34]) for an aggregation procedure (namely, the cumulative exponential weighting procedure) in the model of regression with random design and arbitrary noise distribution having a bounded exponential moment. Thus, one of our objectives in [P10] was to understand whether the ideas used in [30] may be used for obtaining analogous results in the model of regression with deterministic design and non-Gaussian noise.

The idea of aggregating with exponential weights has been discussed by many authors apparently since 1970-ies (see [55] for a nice overview of the subject). Most of the work focused on the important particular case where the set of estimators to aggregate is finite. The inequalities that we proved in [P10] are valid for general set of preliminary estimators satisfying some mild conditions. Furthermore, to treat non-Gaussian errors we introduced new techniques of the proof based on dummy randomization which allowed us to obtain the result for "n-divisible" distributions of errors. We then apply some ideas coming from the Skorokhod embedding [43] to cover the class of all symmetric error distributions with finite exponential moments. Our proofs work in the case when the functions to aggregate are frozen and deterministic. The extension of our results to the case of aggregation of functions depending on data is an interesting open problem.

Finally, as an application, we considered the case where the class $\mathcal{F}$ of functions to aggregate consists of linear combinations of M known functions. As a consequence of our main result we obtained a sparsity oracle inequality (SOI). We refer to [50] where the notion of SOI is introduced in a general context. In an informal way, our result advocates for using as estimator of the coefficients of the unknown "best" linear combination the posterior empirical mean in the model of linear regression with additive Gaussian noise with a sufficiently large variance, even if the noise of the true model is not necessarily Gaussian. In the case when the unknown coefficients of the regression have sparse structure, the use of a prior distribution with density decreasing polynomially at infinity appeared to lead to a nice remainder term in the SOI.

2

Main Results

In this chapter, we briefly present the most important results obtained in papers [P1]-[P10]. Instead of stating the results in whole generality, we will only give their simplest versions. For a discussion on possible extensions, the interested reader is referred to the manuscripts, which can be downloaded from my web page.

2.1 Continuously observed diffusion processes

Let X be a diffusion process given as the solution of the stochastic differential equation

$$dX_t = S(X_t)\,dt + \sigma(X_t)\,dW_t, \quad X_0 = \xi, \quad t \geq 0, \tag{2.1}$$

where W is a standard Brownian motion and the initial value ξ is a random variable independent of W. We assume that a continuous record of observations $X^T = (X_t,\, 0 \leq t \leq T)$ is available. The goal is to estimate the function $S(\cdot)$ or some functional of it. We consider the case of ergodic diffusions: that is X, which is a Markov process, admits an invariant measure.

2.1.1 Sharp adaptive estimation of the drift function

The purpose of the paper [P4] is to propose an estimator of the drift function of a one-dimensional diffusion which is asymptotically minimax up to the optimal constant simultaneously for a large variety of Sobolev balls. Let f_S denote the density with respect to the Lebesgue measure on $\mathbb{R}$ of the invariant measure of the diffusion process defined by (2.1) (cf. [23, Ch. 4, § 18] for more details). To quantify the performance of an estimator

$S_T(\cdot) = S_T(\cdot, X^T)$ of the drift $S(\cdot)$, we use the weighted L^2-risk :

$$R_T(S_T, S) = \int_{\mathbb{R}} \mathbf{E}_S\big[\big(S_T(x) - S(x)\big)^2\big] f_S^2(x)\, dx, \tag{2.2}$$

where $\mathbf{E}_S$ is the expectation with respect to the law $\mathbf{P}_S$ of X defined by (2.1). We call an estimating procedure *adaptive* if its realization does not require any a priori information on the estimated function. The only information that we may (and should) use is the one contained in the observations. We call an estimating procedure *minimax sharp adaptive* or simply *sharp adaptive* over some functional class Σ, if it is adaptive and its "worst case" risk over Σ converges with the best possible rate to the best possible constant.

Let $K(\cdot), Q(\cdot) \in L^2(\mathbb{R})$ be two positive k-times ($k \geq 1$) continuously differentiable symmetric functions such that $\int K = \int Q = 1$, and let $\alpha = \alpha_T$ and $\nu = \nu_T$ be two positive functions of T decreasing to zero as $T \to \infty$. We define the kernel-type estimator of S at the point x by

$$\hat{S}_T(x) = \frac{\frac{1}{\alpha^2}\int_0^T K'\big(\frac{x - X_t}{\alpha}\big)\sigma^2(X_t)\, dt}{\frac{2}{\nu}\int_0^T Q\big(\frac{x - X_t}{\nu}\big)\, dt + \frac{2\varepsilon}{\nu}\, e^{-\ell_T|x|}}, \tag{2.3}$$

where $\varepsilon = \varepsilon_T = e^{\sqrt{\log T}}$ and $\ell_T = (\log T)^{-1}$. One can come to this estimator using the well known formula

$$\big(\sigma^2(x) f_S(x)\big)' = 2S(x) f_S(x). \tag{2.4}$$

In view of the occupations time formula and the martingale representation of the local time, one can check that $(T\alpha^2)^{-1}\int_0^T K'\big((x - X_t)/\alpha\big)\sigma^2(X_t)\, dt$ is a consistent estimator of $\big(\sigma^2(x) f_S(x)\big)'$. Likewise, $2(T\nu)^{-1}\int_0^T Q\big((x - X_t)/\nu\big)\, dt$ is a consistent estimator of $2f_S(x)$. It is now quite natural to define the estimator of $S(x)$ as the quotient of these two estimators.

To simplify the exposition, form now on we suppose that the diffusion coefficient $\sigma(\cdot)$ is identically equal to one. For any function $h \in L^2(\mathbb{R})$, let us denote by $\varphi_h(\cdot)$ the Fourier transform of $h(\cdot)$ defined as $\varphi_h(\lambda) = \int_{\mathbb{R}} e^{i\lambda x} h(x)\, dx$. To avoid the double subscripts, we write φ_f instead of φ_{f_S}. It is proven in [P1, P2], that the estimator (2.3) is asymptotically minimax over a properly chosen Sobolev ball $\Sigma(k, R)$ (k is the order of smoothness and R is the radius) if the kernels and the bandwidths are as follows:

$$\alpha_T^* = \left(\frac{4k}{\pi R T (k+1)(2k+1)}\right)^{\frac{1}{2k+1}}, \quad K^*(x) = \frac{1}{\pi}\int_0^1 (1 - u^{k+\rho_T})\cos(ux)\, du, \tag{2.5}$$

$\nu_T = T^{-1/2}$ and $Q(x)$ is any positive, differentiable, symmetric function with support in $[-1, 1]$ and $\int Q(x)dx = 1$. In Eq. (2.5), we used the notation $\rho_T = 1/\log\log(1 + T)$. The estimator (2.3) defined by such bandwidths and kernels will be denoted by $S_T^*(\cdot)$. Note here that the Fourier transform of the kernel K^* is $\varphi_{K^*}(\lambda) = (1 - |\lambda|^{k+\rho_T})_+$. The exact asymptotic behavior of the maximum over $\Sigma(k, R)$ of the risk of this estimator is $T^{-2k/(2k+1)} P(k, R)$, where $P(k, R)$ is Pinsker's constant [41]. Moreover, the following asymptotic relation holds:

$$R_T(S_T^*, S) \leq \frac{\Delta_T(\alpha, \varphi_{K^*}, |\varphi_f|^2)(1 + o_T(1))}{2\pi T},$$

where $o_T(1)$ is a term tending to zero uniformly in S and the functional Δ_T is defined by

$$\Delta_T(\alpha, h, |\varphi_f|^2) = T\int_{\mathbb{R}} \big|\lambda(1 - h(\alpha\lambda))\varphi_f(\lambda)\big|^2 d\lambda + 4\int_{\mathbb{R}} |h(\alpha\lambda)|^2 d\lambda.$$

Since for known k the optimal kernel is given by (2.5), it is natural to select the adaptive kernel among the functions $\{K_\beta(x) = \pi^{-1}\int_0^1(1-u^\beta)\cos(ux)\,du \mid \beta \geq 1\}$ in a data-driven way. Set

$$h_\beta(\lambda) = (1-|\lambda|^\beta)_+, \quad \hat{\varphi}_T(\lambda) = \frac{1}{T}\int_0^T e^{i\lambda X_t}dt.$$

On the one hand, $|\hat{\varphi}_T(\lambda)|^2 - 4/(T\lambda^2)$ is a good estimate of $|\varphi_f(\lambda)|^2$. On the other hand, the minimization of $\Delta_T(\alpha, h_\beta, |\varphi_f|^2)$ w.r.t. parameters α and β is obviously equivalent to the minimiation of $\Delta_T(\alpha, h_\beta, |\varphi_f|^2) - T\int_{\mathbb{R}}\lambda^2|\varphi_f(\lambda)|^2 d\lambda$. This leads us to defining the functional

$$l_T(h) = T\int_{\mathbb{R}}\lambda^2\big(h^2(\lambda) - 2h(\lambda)\big)\big|\hat{\varphi}_T(\lambda)\big|^2 d\lambda + 8\int_{\mathbb{R}}h(\lambda)d\lambda,$$

obtained by substituting $|\varphi_f(\lambda)|^2$ by $|\hat{\varphi}_T(\lambda)|^2 - 4/(T\lambda^2)$ in the expression $\Delta_T(\alpha, h_\beta, |\varphi_f|^2) - T\int_{\mathbb{R}}\lambda^2|\varphi_f(\lambda)|^2 d\lambda$. Let us define

$$\mathcal{H}_T = \Big\{h : x \mapsto (1-|\alpha_i x|^{\beta_j})_+ \;\Big|\; \alpha_i \in [T^{-1/3}, (\log T)^{-1}],\ j = 1, \dots, \lfloor \log T \rfloor\Big\},$$

with $\alpha_i = (1 + 1/\log T)^{-i}$ and $\beta_j = (1 - j/\log T)^{-1}$, for every $i, j \in \mathbb{N}$.

Form now on, $Q(\cdot)$ is a positive, symmetric, differentiable kernel function supported by $[-1, 1]$ and integrable up to one.

Definition 1. *Let $\tilde{h}$ be a minimizer of $l_T(\cdot)$ over $\mathcal{H}_T$, that is $l_T(\tilde{h}_T) = \min_{h\in\mathcal{H}_T} l_T(h)$ and let*

$$\tilde{K}_T(x) = \frac{1}{2\pi}\int_{\mathbb{R}}\tilde{h}_T(\lambda)\cos(\lambda x)\,d\lambda$$

be the inverse Fourier transform of $\tilde{h}_T$. The adaptive estimator $\hat{S}_T$ of the unknown drift S at any point $x \in \mathbb{R}$ is defined by

$$\hat{S}_T(x) = \frac{\int_0^T \tilde{K}'(x - X_t)\,dt}{2\sqrt{T}\int_0^T Q\big((x-X_t)\sqrt{T}\big)\,dt + 2\sqrt{T}\,e^{-\ell_T|x| + \sqrt{\log T}}},$$

where $\ell_T = 1/\log T$.

Note that the function $\tilde{K}_T(\cdot)$ is differentiable, since $\min_j \beta_j > 1$.

To prove that the estimator $\hat{S}_T(\cdot)$ enjoys nice adaptivity properties, we need some assumptions. Recall that the solution of the stochastic differential equation (2.1) is a strong Markov process. We denote by $P_t(S, x, A)$ the transition probability corresponding to the instant t, that is

$$P_t(S, x, A) = \mathbf{P}_S(X_t \in A \mid X_0 = x), \quad \forall x \in \mathbb{R},\ \forall A \in \mathscr{B}(\mathbb{R}).$$

Here $\mathbf{P}_S$ denotes the probability measure on $\big(C(\mathbb{R}), \mathscr{B}_{C(\mathbb{R})}\big)$ induced by the process (2.1). For every $x \in \mathbb{R}$ and $t \geq 0$ the probability measure $P_t(S, x, \cdot)$ is absolutely continuous with respect to the Lebesgue measure. The corresponding density will be denoted by $p_t(S, x, y)$, so that for any integrable function $g(\cdot)$, we have

$$\mathbf{E}_S[g(X_t)|\mathscr{F}_s] = \int_{\mathbb{R}} g(y)\,p_{t-s}(S, X_s, y)\,dy.$$

Let $k > 0$ be an integer and denote by $\Sigma(k)$ the set of all functions satisfying conditions:

C1. S is k-times continuously differentiable in $\mathbb{R}$ and $\overline{\lim}_{|x|\to\infty} \operatorname{sgn}(x)S(x) < 0$.

C2. There exist positive numbers C and ν such that $\left|S^{(k)}(x)\right| \le C\big(1+|x|^{\nu}\big)$, $\forall x \in \mathbb{R}$.

The problem we consider is the following: we know that $\boldsymbol{x}^T$ is a sample path of the process X^T given by (2.1) with a drift function $S \in \Sigma = \cup_{k\ge 1}\Sigma(k)$ and we want to estimate the function $S(\cdot)$. For obtaining minimax results we consider the local setting. For any function $S_0 \in \Sigma(k)$ and for all $\delta > 0$, we define the vicinities

$$\widetilde{V}_\delta(S_0,k) = \Big\{ S \in \Sigma(k) \,\Big|\, \sup_{x\in\mathbb{R}} \left|S^{(i)}(x) - S_0^{(i)}(x)\right| \le \delta, \quad i = 0,1,\dots,k-1 \Big\}.$$

The center of localization $S_0(\cdot)$ is assumed to fulfill the following additional assumptions:

C3. There exist $\kappa > 0$ and $q > 1$ such that $\mathbf{E}_{S_0}\Big[\sup_{y\in\mathbb{R}} p_\kappa^q(S_0, X_0, y)\Big] < \infty$.

C4. Let $\varphi_0(\cdot)$ be the Fourier transform of the invariant density $f_{S_0}(\cdot)$. There exists $\tau > 0$ such that $\int_{\mathbb{R}} |\lambda|^{2k+2+\tau}\left|\varphi_0(\lambda)\right|^2 d\lambda < \infty$.

We define now the Sobolev balls; in our setup they also are weighted by the square of the invariant density. Let us denote

$$\widetilde{\Sigma}_\delta(k,R,S_0) = \Big\{ S \in \widetilde{V}_\delta(S_0,k) \,\Big|\, \int_{\mathbb{R}} \big[(S-S_0)^{(k)}(x)\big]^2 f_S^2(x)\,dx \le R \Big\}.$$

Theorem 1. *Let S_0 satisfy assumptions C1–C4 and let the risk $R_T(\cdot,\cdot)$ be defined by (2.2). If the initial condition ξ follows the invariant law, then*

$$\overline{\lim_{\delta\to 0}}\,\overline{\lim_{T\to\infty}} \sup_{S\in\widetilde{\Sigma}_\delta(k,R,S_0)} T^{\frac{2k}{2k+1}} R_T(\hat{S}_T, S) = P(k,R),$$

where $P(k,R) = (2k+1)\big(\frac{k}{\pi(k+1)(2k+1)}\big)^{\frac{2k}{2k+1}} R^{\frac{1}{2k+1}}$ is Pinsker's constant.

2.1.2 Second-order minimax estimation of the invariant density

Let us switch now our attention to the problem of estimating the invariant density $f_S(\cdot)$. To simplify the computations we assume that $\sigma(x) \equiv 1$, so the invariant density is given by

$$f_S(x) = G(S)^{-1} \exp\big\{2\int_0^x S(v)\,\mathrm{d}v\big\},$$

where $G(S)$ is the normalizing constant. Furthermore, we assume that $S \in \Sigma_{\gamma_*}(A_*, C_*, \nu_*)$, where

$$\Sigma_{\gamma_*}(A_*, C_*, \nu_*) = \left\{ S(\cdot) : \begin{array}{ll} \operatorname{sgn}(x)S(x) \le -\gamma_* & \forall |x| > A_* \\ \left|S(x)\right| \le C_*(1+|x|^{\nu_*}), & \forall x \in \mathbb{R} \end{array} \right\}.$$

Here γ_*, A_*, C_* and ν_* are some (unknown) positive constants.

Fix some integer $k \geq 2$. The function $S(\cdot)$ is supposed to be $(k-2)$-times differentiable with absolutely continuous $(k-2)^{\text{th}}$ derivative and to belong to the set

$$\Sigma(k,R) = \Big\{ S(\cdot) \in \Sigma \,:\, \int_{\mathbb{R}} [f_S^{(k)}(x) - f_{S_*}^{(k)}(x)]^2 \,\mathrm{d}x \leq R \Big\},$$

where $R > 0$ is some constant and $f_S^{(k)}(\cdot)$ is the k-th derivative (in the distributional sense) of the function $f_S(\cdot)$. The set $\Sigma(k,R)$ is a Sobolev ball of smoothness k and radius R centered at $f_{S_*} = f_*$. The choice of the center is not arbitrary, it is assumed to be smoother than the other functions of the class. For simplicity, we focus our attention on the case $S_*(x) = -x$ corresponding to an Ornstein-Uhlenbek process. Finally we define the parameter set $\Sigma_* = \Sigma_*(k,R) = \Sigma(k,R) \cap \Sigma_{\gamma_*}(A_*, C_*, \nu_*)$.

In this setting, the problem of first-order minimax estimation of $f_S(\cdot)$ under mean integrated squared loss has been studied by Kutoyants [31], who proved that the minimax rate of estimation is $T^{-1/2}$ and the nonparametric analogue of the Fisher information is given by

$$\mathrm{I}(S,x) = \left[4 f_S(x)^2 \, \mathbf{E}_S \left(\frac{\chi_{\{\xi > x\}} - F_S(\xi)}{f_S(\xi)} \right)^2 \right]^{-1},$$

where ξ is supposed to follow the invariant law and $F_S(\cdot)$ is the c.d.f. associated to the probability density $f_S(\cdot)$. Moreover, it is shown that under mild regularity conditions the local-time estimator

$$f_T^\circ(x) = \frac{1}{T} \int_0^T \mathrm{sgn}(x - X_t) \,\mathrm{d}X_t + \frac{|X_T - x| - |X_0 - x|}{T},$$

kernel-type estimators $\bar{f}_{K,T}(x)$ and a wide class of unbiased estimators $\tilde{f}_T(x)$ are consistent, asymptotically normal and asymptotically (first-order) minimax.

In order to discriminate between these first-order minimax estimators, we propose to study the second-order risk

$$\mathscr{R}_T(\bar{f}_T, f_S) = \int_{\mathbb{R}} \mathbf{E}_S[(\bar{f}_T(x) - f_S(x))^2] \mathrm{d}x - T^{-1} \int_{\mathbb{R}} I(S,x)^{-1} dx,$$

where $\bar{f}_T(x)$ is an arbitrary estimator of the density. It is evident that for first-order asymptotically minimax estimators $\bar{f}_T$, the quantity $T\mathscr{R}_T(\bar{f}_T, f_S)$ tends to zero uniformly in $\Sigma_*(k,R)$, as $T \to \infty$. It can be shown that for some of these estimators there exists a non degenerate limit for $T^{\frac{2k}{2k-1}} \sup_{S \in \Sigma_*(k,R)} \mathscr{R}_T(\bar{f}_T, f_S)$ and for the others this limit is equal to infinity. Therefore we can compare the performance of these estimators according to the limits of this quantity. The following result describes what is its lowest possible limiting value.

Theorem 2. *For every integer $k \geq 2$ and for every $R, \gamma_*, A_*, C_*, \nu_* > 0$, it holds*

$$\lim_{T \to \infty} \Big\{ \inf_{\bar{f}_T} \sup_{S \in \Sigma_*} T^{\frac{2k}{2k-1}} \mathscr{R}_T(\bar{f}_T, f_S) \Big\} = -\widehat{P}(k,R),$$

where $\widehat{P}(k,R) = 2(2k-1)\big(\frac{4k}{\pi(k-1)(2k-1)}\big)^{2k/(2k-1)} R^{-1/(2k-1)}$.

It is noteworthy that the estimator proved to achieve the minimax bound stated in Theorem 2 is independent on γ_*, A_*, C_* and ν_*, but relies on the knowledge of parameters k and R.

We should also acknowledge that the estimator proposed in [P3] is rather complicated for computation and in most cases it would be better to use simpler estimator which is not necessarily second-order efficient up to the constant. For example, the kernel estimator with properly chosen bandwidth is second-order rate-minimax and is easier to compute that the estimator proposed in [P3].

2.1.3 Statistical equivalence for scalar ergodic diffusions

Let us briefly introduce some basic notation such that we can announce the main results. For some fixed constants C, A, $\gamma > 0$ we consider the nonparametric drift class

$$\Sigma \triangleq \Big\{ S \in \mathrm{Lip}_{\mathrm{loc}}(\mathbb{R}) : \sup_{x\in\mathbb{R}} |S(x)|/(1+|x|) \le C,\ \sup_{|x|\ge A} S(x)\,\mathrm{sgn}(x) \le -\gamma \Big\}, \tag{2.6}$$

where $\mathrm{Lip}_{\mathrm{loc}}(\mathbb{R})$ denotes the set of locally Lipschitz continuous functions $S : \mathbb{R} \to \mathbb{R}$ and $\mathrm{sgn}(x) \triangleq x/|x|$. For a drift $S_0 \in \Sigma$ and for any density $f_0 \in L^1(\mathbb{R})$ we introduce their local neighborhood with parameters ε, $\zeta, \eta > 0$

$$\Sigma_{\varepsilon,\eta,\zeta}(S_0, f_0) = \Big\{ S \in \Sigma : \|(S-S_0)^2\sqrt{f_S}\|_1 \le \varepsilon^2, \|(S-S_0)^2(f_S - f_0)\|_1 \le \eta^2, \|f_S - f_0\|_1 \le \zeta \Big\}.$$

It is natural to consider neighborhoods around (S_0, f_{S_0}), but it is by no means necessary for the calculations to enforce $f_0 = f_{S_0}$.

We now define precisely the local experiments $\mathbb{E}_1$ and $\mathbb{F}_1$, for which we shall prove asymptotic equivalence. Note that we define the Gaussian shift experiment on the space $\mathbb{R}^{L^2(\mathbb{R})}$ and not on $C(\mathbb{R})$ via the natural interpretation of the differentials as integrators for $L^2(\mathbb{R})$-functions. Of course, the law is already characterized by the integration of the functions $\mathbf{1}_{[0,y]}$, $y \in \mathbb{R}$, which corresponds to the signal in white noise interpretation on the space $C(\mathbb{R})$ up to the knowledge of the value at zero.

Definition 2. *We define the diffusion experiment localized around* (S_0, f_0)

$$\mathbb{E}_1 \triangleq \mathbb{E}_1(S_0, f_0, T, \varepsilon, \eta, \zeta) \triangleq (C([0,T]), \mathcal{B}_{C([0,T])}, (\mathbf{P}_S^T)_{S\in\Sigma_{\varepsilon,\eta,\zeta}(S_0,f_0)}),$$

$\mathbf{P}_S^T$ *being the law of the stationary diffusion process with drift* S *on the canonical space* $C([0,T])$. *The Gaussian shift experiment localized around* (S_0, f_0) *is given by*

$$\mathbb{F}_1 \triangleq \mathbb{F}_1(S_0, f_0, T, \varepsilon, \eta, \zeta) \triangleq (\mathbb{R}^{L^2(\mathbb{R})}, \mathcal{B}_{\mathbb{R}}^{\otimes L^2(\mathbb{R})}, (\mathbf{Q}_S^T)_{S\in\Sigma_{\varepsilon,\eta,\zeta}(S_0,f_0)}),$$

where $\mathbf{Q}_S^T$ *denotes the law of the Gaussian shift experiment*

$$dZ_x = S(x) f_0(x)^{1/2}\,dx + T^{-1/2}\,dB_x, \quad x \in \mathbb{R},$$

with a Brownian motion B *on the real line.*

Let $\Delta(\mathbb{E}, \mathbb{F})$ be the Le Cam pseudo-distance between arbitrary two experiments $\mathbb{E}$ and $\mathbb{F}$ (see [33] for the precise definition).

Theorem 3. *If for $T \to \infty$ the asymptotics $\varepsilon_T = o(T^{-1/4})$, $\eta_T = o(T^{-1/2})$ and $\zeta_T = o(1)$ hold, then the following convergence holds true uniformly in $S_0 \in \Sigma$:*

$$\lim_{T\to\infty} \Delta\Big(\mathbb{E}_1\big(S_0, f_0, T, \varepsilon_T, \eta_T, \zeta_T\big), \mathbb{F}_1\big(S_0, f_0, T, \varepsilon_T, \eta_T, \zeta_T\big)\Big) = 0.$$

Without going into details, I would like to say some words about the proof of this theorem. The only thing we need to know about the Le Cam's distance is that

P1 If the experiments $\mathbb{E}$ and $\mathbb{F}$ have the same parameter space Θ and are dominated, then the equality in law of likelihood processes (indexed by $\vartheta \in \Theta$) of these experiments entails their equivalence, that is $\Delta(\mathbb{E}, \mathbb{F}) = 0$.

P2 If the experiments $\mathbb{E}$ and $\mathbb{F}$ are defined on the same probability space, have the same parameter space Θ and are dominated, then the Le Cam distance between $\mathbb{E}$ and $\mathbb{F}$ is upper bounded up to a multiplicative constant by the supremum in ϑ of the Kullback-Leibler divergence between the likelihoods of $\mathbb{E}$ and $\mathbb{F}$.

Using the Girsanov and the occupation time formulas, the likelihood of the diffusion experiment can be written as

$$\mathcal{L}_T(S) = \exp\Big\{ \int_0^T (S - S_0)(X_t)\, dW_t - \frac{1}{2}\int_{\mathbb{R}} (S - S_0)^2(x) L_T^x(X)\, dx \Big\},$$

where $L_T^x(X)$ is the local time of the diffusion process X at the point $x \in \mathbb{R}$ up to time $T \geq 0$.

Let us now introduce two auxiliary experiments $\mathbb{E}_2$ and $\mathbb{F}_2$. We define the local experiment $\mathbb{E}_2 = \mathbb{E}_2(S_0, f_0, T, \varepsilon, \eta, \zeta)$ by the observations (X^T, V), where X^T is a path of ergodic diffusion with drift S over $[0, T]$ and $V = (V_x,\, x \in \mathbb{R})$ is given by

$$dV_x = S_0(x)(Tf_0(x) - L_T^x(X))_+^{1/2}\, dx + dB_x, \quad x \in \mathbb{R}, \tag{2.7}$$

where B stands for a two-sided Brownian motion on $\mathbb{R}$ independent of W and X_0 and $A_+ = \max(A, 0)$. The second auxiliary experiment, denoted by $\mathbb{F}_2 = \mathbb{F}_2(S_0, f_0, T, \varepsilon, \eta, \zeta)$, is defined by observing the pair (Y, V), where Y is a weak solution of the SDE

$$dY_t = \big(S(Y_t)\mathbf{1}_{\{L_t^{Y_t}(Y) \leq Tf_0(Y_t)\}} + S_0(Y_t)\mathbf{1}_{\{L_t^{Y_t}(Y) > Tf_0(Y_t)\}}\big)\, dt + dW_t, \quad t \in [0, T],$$

with initial distribution $Y_0 \sim f_0$, and V is the conditionally to Y Gaussian process

$$dV_x = S(x)(Tf_0(x) - L_T^x(Y))_+^{1/2}\, dx + dB_x, \quad x \in \mathbb{R}, \tag{2.8}$$

where B is the same as in (2.7).

We proved in [P5] that $\Delta(\mathbb{E}_1, \mathbb{E}_2) = \Delta(\mathbb{F}_1, \mathbb{F}_2) = 0$. Note that the first equality is easily understandable. Indeed, if we have at our disposal the observation X^T, we gain no information about S by observing V from (2.7). So inference in $\mathbb{E}_2$ is not easier than in $\mathbb{E}_1$. On the other hand, since the observation in $\mathbb{E}_2$ comprises the observation in $\mathbb{E}_1$, the inference in $\mathbb{E}_1$ is not easier than the inference in $\mathbb{E}_2$. So it is not surprising that these experiments are statistically equivalent.

The equivalence of $\mathbb{F}_1$ and $\mathbb{F}_2$ is far less obvious. It is proved using the aforementioned property **P1** of the Le Cam distance. Indeed, one easily checks that the log-likelihood of the model having as observation (Y, V) is given by $\mathcal{Z}(S) - \frac{1}{2}\langle \mathcal{Z}(S) \rangle$, where

$$\mathcal{Z}(S) = \int_0^T (S - S_0)(Y_t) \mathbf{1}_{\{L_t^{Y_t}(Y) \le T f_0(Y_t)\}} \, dW_t + \int_{\mathbb{R}} (S - S_0)(x) (T f_0(x) - L_T^x(Y))^{1/2} \, dB_x,$$
$$\langle \mathcal{Z}(S) \rangle = \int_0^T (S - S_0)^2(Y_t) \mathbf{1}_{\{L_t^{Y_t}(Y) \le T f_0(Y_t)\}} \, dt + \int_{\mathbb{R}} (S - S_0)^2(x) (T f_0(x) - L_T^x(Y)) \, dx.$$

Using the extended occupation time formula [43, Ex. VI.1.15], one checks that $\langle \mathcal{Z}(S) \rangle = T \int_{\mathbb{R}} (S - S_0)^2 f_0$. This entails that $\mathcal{Z}(S)$ is a Gaussian random variable (this can be checked by computing its Laplace transform). Since $\mathcal{Z}(S)$ is a linear functional of $S - S_0$, the same argument implies that $\alpha_1 \mathcal{Z}(S_1) + \ldots + \alpha_p \mathcal{Z}(S_p)$ is a Gaussian random variable for every $\alpha = (\alpha_1, \ldots, \alpha_p) \in \mathbb{R}^p$ and $S_1, \ldots, S_p \in \Sigma$. Therefore, the log-likelihood of the experiment $\mathbb{F}_2$ is a Gaussian process having the same mean and covariance operator as the log-likelihood of the experiment $\mathbb{F}_1$. This implies that $\Delta(\mathbb{F}_1, \mathbb{F}_2) = 0$.

To complete the proof of Theorem 3, it suffices to compute the Kullback-Leibler divergence between the likelihoods of the experiments $\mathbb{E}_2$ and $\mathbb{F}_2$ (which are defined on the same measurable space) and to show that it goes to zero as T tends to infinity.

It is possible to give a number of statistical experiments equivalent to $\mathbb{F}_1$.

Remark 1. *The following experiments are equivalent to $\mathbb{F}_1$ and $\mathbb{F}_2$ for parameters $S \in \Sigma_{\varepsilon,\eta}(S_0)$:*

$$dY_x = S(x)\, dx + T^{-1/2} f_0(x)^{-1/2} \, dB_x, \quad x \in \mathbb{R},$$
$$dY_x = (S(x) - S_0(x)) \sqrt{f_0(x)}\, dx + T^{-1/2} \, dB_x, \quad x \in \mathbb{R},$$
$$dY_x = b(F_{f_0}^{-1}(x))\, dx + T^{-1/2} \, dB_x, \quad x \in (0,1),$$

where $F_f(x) = \int_{-\infty}^x f(y)\, dy$ and dB is Gaussian white noise. For the proof it suffices to check that the laws of the likelihood processes coincide.

Remark 2. *The preceding asymptotic equivalence result holds in particular for the local parameter subclass $\tilde{\Sigma}_{\varepsilon,T}(S_0, f_0) \triangleq \left\{ S \in \Sigma \,\middle|\, \|(S - S_0)^2 \sqrt{f_S}\|_1 \le \varepsilon^2, \, \|f_S^{1/2} - f_0 f_S^{-1/2}\|_\infty \le T^{-1/2} \right\}$, when $\varepsilon = \varepsilon_T = o(T^{-1/4})$ for $T \to \infty$.*

2.1.4 Statistical equivalence for multidimensional ergodic diffusions

We now address the issu of extending previous results to the case of multidimensional ergodic diffusions. Note that on the one hand, even for simple experiments, as the classical ones described above, results for asymptotic equivalence in the multidimensional case are very scarce. We only know of the recent work by Carter [10] who proved asymptotic equivalence for two-dimensional Gaussian regression, but argued that his method fails for higher dimensions. Brown and Zhang [9] remarked that the two classical experiments and their accompanying Gaussian shift experiments are not asymptotically equivalent in the case of nonparametric classes of Hölder regularity $\beta \le d/2$, where d denotes the dimension. In the

recent work [42] Reiss proved that the statistical equivalence between the regression experiment and signal in Gaussian white noise model holds for $\beta > d/2$.

On the other hand, the methodology we used in the previous section to establish asymptotic equivalence for scalar diffusions relied heavily on the concept of local time. For multidimensional diffusions local time does not exist. This might explain why the statistical theory for scalar diffusions is very well developed [32], while inference problems for multidimensional diffusions are more involved and much less studied. We refer to Bandi and Moloche [4] for the analysis of kernel estimators for the drift vector and the diffusion matrix and to Aït-Sahalia [2] for a recent discussion of applications for multidimensional diffusion processes in econometrics.

We assume that a continuous record $X^T = \{X_t,\ 0 \le t \le T\}$ of a d-dimensional diffusion process X is observed up to time instant T. We denote by $S_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \dots, d$, the components of the vector valued function S. In what follows, we assume that the drift is of the form $S = -\nabla V$, where $V \in C^2(\mathbb{R}^d)$ is referred to as potential, and $\sigma \equiv I_d$. This restriction permits to use strong analytical results for the Markov semigroup of the diffusion on the L^2-space generated by the invariant measure.

For positive constants M_1 and M_2, we define $\Sigma(M_1, M_2)$ as the set of all functions $S = -\nabla V : \mathbb{R}^d \to \mathbb{R}^d$ satisfying for any $x, y \in \mathbb{R}^d$

$$|S(x)| \le M_1(1 + |x|), \tag{2.9}$$

$$(S(x) - S(y))^T(x - y) \le -M_2|x - y|^2, \tag{2.10}$$

where $|\cdot|$ denotes the Euclidian norm in $\mathbb{R}^d$. Any such function S is locally Lipschitz-continuous. Therefore equation (2.1) has a unique strong solution, which is a homogeneous continuous Markov process, cf. [45, Thm. 12.1]. Set $G(S) = \int_{\mathbb{R}^d} e^{-2V(u)}\, du$ and

$$f_S(x) = G(S)^{-1} e^{-2V(x)}, \qquad x \in \mathbb{R}^d.$$

Under condition (2.10) we have $G(S) < \infty$ and the process X is ergodic with unique invariant probability measure [6, Thm. 3.5]. Moreover, the invariant probability measure of X is absolutely continuous with respect to the Lebesgue measure and its density is f_S. From now on, we assume that the initial value ξ in (2.1) follows the invariant law such that the process X is strictly stationary.

We write $f_S(h) \triangleq \mathbf{E}_S[h(X_0)] = \int h\, f_S$. Let $P_{S,t}$ be the transition semigroup of this process on $L^2(f_S)$, that is

$$P_{S,t}h(x) = \mathbf{E}_S[h(X_t)|X_0 = x],\ h \in L^2(f_S).$$

The transition density is denoted by $p_{S,t}$: $P_{S,t}f(x) = \int f(y)p_{S,t}(x, y)\, dy$.

For any multi-index $\alpha \in \mathbb{N}^d$ and $x \in \mathbb{R}^d$ we set $|\alpha| = \alpha_1 + \ldots + \alpha_d$ and $x^\alpha = x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d}$. Let us introduce the Hölder class

$$\mathcal{H}(\beta, L) = \left\{ f \in C^{\lfloor\beta\rfloor}(\mathbb{R}^d; \mathbb{R}) :\ \begin{array}{c} |D^\alpha f(x) - D^\alpha f(y)| \le L|x - y|^{\beta - \lfloor\beta\rfloor} \\ \text{for any } \alpha \text{ such that } |\alpha| = \lfloor\beta\rfloor \end{array} \right\}$$

where $\lfloor\beta\rfloor$ is the largest integer *strictly* smaller than β and $D^\alpha f \triangleq \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$.

Let $\Sigma_\beta(L, M_1, M_2)$ be the set of functions $S \in \Sigma(M_1, M_2)$ such that all d components S_i of S are in $\mathcal{H}(\beta, L)$. We fix a function $S^\circ \in \Sigma_\beta(L, M_1, M_2)$. Suppose $\Sigma \subset \Sigma(M_1, M_2)$ for some $M_1, M_2 > 0$. For any $T > 0$ let $\mathbb{E}(\Sigma, T)$ be the statistical experiment of observing the diffusion with $S \in \Sigma$.

For any function $S \in L^2(f_{S^\circ}; \mathbb{R}^d) = \{h : \mathbb{R}^d \to \mathbb{R}^d : \int |h|^2 f_{S^\circ} < \infty\}$ we denote by $\mathbf{Q}_{S,T}$ the Gaussian measure on $(C(\mathbb{R}^d; \mathbb{R}^d), \mathcal{B}_{C(\mathbb{R}^d;\mathbb{R}^d)})$ induced by the d-dimensional process Z satisfying

$$dZ(x) = S(x)\, f_{S^\circ}(x)^{1/2} dx + T^{-1/2}\, dB(x), \qquad Z(\mathbf{0}) = \mathbf{0}, \qquad x \in \mathbb{R}^d, \tag{2.11}$$

where $B(x) = (B_1(x), \dots, B_d(x))$ and $B_1(x), \dots, B_d(x)$ are independent d-variate Brownian sheets, that is zero mean Gaussian processes with $\mathrm{Cov}(B_i(x), B_i(y)) = |R_x \cap R_y|$ where $R_x = \{u \in \mathbb{R}^d : u_i \in [0, x_i]\}$.

Definition 3 (Gaussian shift experiment). *For $\Sigma \subset L^2(f_{S^\circ}; \mathbb{R}^d)$ and $T > 0$ let $\mathbb{F}(\Sigma, T)$ be the Gaussian shift experiment (2.11) with $S \in \Sigma$, that is $\mathbb{F}(\Sigma, T) = \big(C(\mathbb{R}^d; \mathbb{R}^d), \mathcal{B}_{C(\mathbb{R}^d;\mathbb{R}^d)}, (\mathbf{Q}_{S,T})_{S\in\Sigma}\big)$.*

For any positive numbers ε, η and for any hypercube $A \subset \mathbb{R}^d$, we define the local neighborhood of S°

$$\Sigma(S^\circ, \varepsilon, \eta, A) = \left\{ S \in \Sigma_\beta(L, M_1, M_2) : \begin{matrix} |S(x) - S^\circ(x)| \le \varepsilon \mathbf{1}_A(x),\ x \in \mathbb{R}^d, \\ |f_S(x) - f_{S^\circ}(x)| \le \eta f_{S^\circ}(x),\ x \in A \end{matrix} \right\},$$

where $\mathbf{1}_A$ is the indicator function of the set A.

Theorem 4. *If ε_T and η_T satisfy the conditions*

$$\lim_{T\to\infty} T^{-\beta} \varepsilon_T^{2-d} = \lim_{T\to\infty} T^{\frac{1}{4}+\frac{d-2}{8\beta}} \varepsilon_T (\log(T\varepsilon_T^{-1}))^{\mathbf{1}(d=2)} = \lim_{T\to\infty} T\eta_T\varepsilon_T^2 = 0,$$

then the multidimensional diffusion model is asymptotically equivalent to the Gaussian shift model (2.11) over the parameter set $\Sigma_{0,T} = \Sigma(S^\circ, \varepsilon_T, \eta_T, A)$, that is

$$\lim_{T\to\infty} \sup_{S^\circ \in \Sigma_\beta(L,M_1,M_2)} \Delta\big(\mathbb{E}(\Sigma_{0,T}, T), \mathbb{F}(\Sigma_{0,T}, T)\big) = 0.$$

Let us see for which Hölder regularity β on the drift an estimator can attain the local neighborhood, that is $|\hat{S}_{h(T),T}(x) - S(x)| \le \varepsilon_T$ and $|\hat{f}_{h(T),T}(x) - f_S(x)| \le \eta_T$ hold with a probability tending to one (cf. Nussbaum [39] for this concept). By the rates obtained in [P6, Corollary 1] (see also [7]), and the condition in Theorem 4, this is the case if $\beta > (d - 1 + \sqrt{2(d-1)^2 - 1})/2$. The critical regularity thus grows like $(1/2 + 1/\sqrt{2})d$ for $d \to \infty$. In dimension 2 we obtain the condition $\beta > 1$ as in the result by Carter [10] for Gaussian regression. Whether for Hölder classes of smaller regularity asymptotic equivalence fails, remains a challenging open problem.

2.2 Second-order efficiency in semiparametrics

Consider the "signal in Gaussian white noise model", that is the observations $(x^\varepsilon(t),\ t \in [-1/2, 1/2])$ with

$$dx^\varepsilon(t) = f_\vartheta(t)\, dt + \varepsilon dW(t), \qquad t \in [-1/2, 1/2], \tag{2.12}$$

are available, where $W(t)$ is a Brownian motion. Assume that the signal has the form $f_\vartheta(t) = f(t-\vartheta)$, where $f : \mathbb{R} \to \mathbb{R}$ is a symmetric periodic function having 1 as smallest period. More precisely, we assume that the function f belongs to the set $\mathcal{F}_0 = \cup_{\rho>0}\mathcal{F}_0(\rho)$ with

$$\mathcal{F}_0(\rho) = \Big\{ f \in L^2_{loc} \,:\, f(x) = f(-x) = f(x+1),\ \forall\, x \in \mathbb{R};\ |f_1| \geq \rho \Big\},$$

where we denote by L^2_{loc} is the set of all locally squared integrable functions and by $f_1 = \sqrt{2}\int_{-1/2}^{1/2} f(t)\,\cos(2\pi t)\,dt$.

The goal is to estimate the parameter ϑ, which is assumed to lie in $\Theta \subset]-T, T]$ with $T < 1/4$. As explained in [P7], the assumption $T < 1/4$ is necessary for the identifiability of the parameter ϑ. In this context, the unknown function f is considered as an infinite dimensional nuisance parameter. The Fisher information in the problem of estimating ϑ with fixed f is

$$I^\varepsilon(f) = \varepsilon^{-2}\int_{-1/2}^{1/2} f'(x)^2\,dx = \varepsilon^{-2}\sum_{k\in\mathbb{N}} (2\pi k)^2 f_k^2,$$

where $f_k = \sqrt{2}\int_{-1/2}^{1/2}\cos(2\pi k t)\,f(t)\,dt \triangleq \varepsilon^{-2}\|f'\|^2$.

We call *filtering sequence* or *filter* any $h = (h_k)_{k\in\mathbb{N}} \in [0,1]^{\mathbb{N}}$ such that only a finite number of coefficients h_k are non-zero. Define the functional

$$\Phi_\varepsilon(\tau, h) \triangleq \sum_{k=1}^{\infty} h_k \bigg(\int_{-1/2}^{1/2} \cos[2\pi k(t-\tau)]\,dx^\varepsilon(t) \bigg)^2. \tag{2.13}$$

It is easy ti check that the Penalized Maximum Likelihood Estimator (PMLE) of ϑ is then $\hat\vartheta_{\text{PMLE}} \triangleq \arg\max_\tau \Phi_\varepsilon(\tau, h)$. The role of the sequence h is to filter out the irrelevant terms in the right side of (2.13). That is, for a "nice" filter h the values h_k corresponding to a small signal-to-noise ratio $|f_k|/\varepsilon$ are close to zero.

For deterministic filters h, the asymptotic behavior of the estimator $\hat\vartheta_{\text{PMLE}}$ is studied in [P7]. Under some smoothness assumptions on f, for a broad choice of filters h, $\hat\vartheta_{\text{PMLE}}$ is proved to be first-order asymptotically efficient. Moreover, it is shown that the second-order term of its risk expansion is $\varepsilon^2 R^\varepsilon[f,h]/\|f'\|^4$, where

$$R^\varepsilon[f,h] \triangleq \sum_{k=1}^{\infty} (2\pi k)^2 \big[(1-h_k)^2 f_k^2 + \varepsilon^2 h_k^2\big].$$

This result suggests to use the filter $h_{opt} = \arg\min_h R^\varepsilon[f,h]$ for defining the PMLE of ϑ. However, this minimizer is uncomputable in practice since it depends on f. To get rid of this dependence, the minimax approach recommends the utilization of the filter $h_{\mathcal{F}} = \arg\min_h \sup_{f\in\mathcal{F}} R^\varepsilon[f,h]$. If $\mathcal{F}$ is a ball in a Sobolev space, a solution of this minimization problem is given by the Pinsker filter. The properties of the estimator based on this filter are studied in [P7, Thm. 2 and 3]. Here we will state the results concerning the adaptive choice of the filtering sequence and the quality of the resulting estimator of ϑ.

To define the estimator, we need the notation

$$\begin{aligned} x_k &= \sqrt{2}\int_{-1/2}^{1/2}\cos(2\pi k t)\,dx^\varepsilon(t), \\ x_k^* &= \sqrt{2}\int_{-1/2}^{1/2}\sin(2\pi k t)\,dx^\varepsilon(t). \end{aligned} \tag{2.14}$$

The adaptive procedure:

1. Choose $\beta_* > 1$ and set $N_\varepsilon = 5 \vee [(\varepsilon^2 \log \varepsilon^{-5})^{-\frac{1}{2\beta_*+1}}]$, $\nu_\varepsilon = [e^{\sqrt{\log N_\varepsilon}}]$ and $\rho_\varepsilon = \nu_\varepsilon^{-1/3}$.

2. Define the sequence $(\kappa_j)_{j\geq 1}$ by

$$\kappa_j = \begin{cases} (1+\nu_\varepsilon)^{j-1}, & j = 1,2, \\ \kappa_{j-1} + \lfloor \nu_\varepsilon \rho_\varepsilon (1+\rho_\varepsilon)^{j-2} \rfloor, & j = 3,4,\dots, \end{cases} \tag{2.15}$$

   and the blocks $B_j = \{k \in \mathbb{N} : \kappa_j \leq k < \kappa_{j+1}\}$.

3. Set $\varphi_j = \sqrt{24 \log \varepsilon^{-5}/(\kappa_{j+1} - \kappa_j)}$, $\sigma_j^2 = \sum_{B_j} (2\pi k)^2$ and define

$$\hat{h}_k^S = \left(1 - \frac{\varepsilon^2 \sigma_j^2 (1+\varphi_j)}{(\|y'\|_{(j)}^2 - 2\varepsilon^2\sigma_j^2)_+ + \varepsilon^2 \sigma_j^2}\right)_+ \mathbf{1}_{\{j \leq N_\varepsilon\}}, \quad \forall k \in B_j \tag{2.16}$$

   with $\|y'\|_{(j)}^2 = \sum_{k\in B_j} (2\pi k)^2 |y_k|^2$, $y_k = x_k + i x_k^*$

4. Compute the preliminary estimator $\bar{\vartheta}_\varepsilon$ by

$$\bar{\vartheta}_\varepsilon = \begin{cases} \frac{1}{2\pi} \arctan\left(x_1^*/x_1\right), & x_1 \neq 0, \\ 1/4, & x_1 = 0. \end{cases}$$

5. Define $\hat{\vartheta}_\varepsilon^S$ as the minimum in $\bar{\Theta}_\varepsilon = [\bar{\vartheta}_\varepsilon - \varepsilon \log(\varepsilon^{-2}), \bar{\vartheta}_\varepsilon + \varepsilon \log(\varepsilon^{-2})]$ of $\Phi_\varepsilon(\cdot, \hat{h}^S)$ (see (2.13)).

Note that the only "free" parameter in this procedure is β_*. In practice, if no information on the regularity of f is available, it appears plausible to assume that f has Sobolev smoothness $\beta_* = 2$.

Let T_j be the length of the block B_j and $T_\varepsilon = \inf_j T_j$. The oracle choice of h in the class $\mathcal{H}^*(B)$ of all filters constant on the blocks $B = \{B_j\}_j$ is denoted by h^*: $R^\varepsilon[f, h^*] = \min_{h\in\mathcal{H}^*(B)} R^\varepsilon[f,h]$. Introduce the functional class

$$\mathcal{F}(\beta_*, L_*, \rho) = \left\{ f \in \mathcal{F}_0(\rho) : \|f^{(\beta_*)}\| \leq L_* \right\},$$

where $\beta_* > 1$, $\rho > 0$, $L_* > 0$ are some constants.

Theorem 5 (Oracle inequality). *If the blocks B_j verify $\log \varepsilon^{-1} = o(T_\varepsilon)$ as $\varepsilon \to 0$, then*

$$I^\varepsilon(f) \mathbf{E}_{\vartheta,f}[(\hat{\vartheta}_\varepsilon^S - \vartheta)^2] \leq 1 + (1+\alpha_\varepsilon) \frac{R^\varepsilon[f, h^*]}{\|f'\|^2},$$

where $\alpha_\varepsilon \to 0$ as $\varepsilon \to 0$ uniformly in $f \in \mathcal{F}(\beta_*, L_*, \rho)$.

Remark 3. *If the block B_j is large, then more observations (x_k, x_k^*) are used for estimating the value of the oracle $h^*_{\kappa_j}$. Hence, it is natural to expect that α_ε decreases as T_ε increases. A thorough inspection of the proof allows to describe this feature with the help of the order relation $\alpha_\varepsilon^2 \asymp T_\varepsilon^{-1} \log \varepsilon^{-1}$.*

Now we consider the class $\mathcal{H}_{\mathrm{mon}}$ of filters having decreasing components, that is

$$\mathcal{H}_{\mathrm{mon}} = \left\{ h \in [0,1]^{\mathbb{N}} \ : \ h_k \geq h_{k+1}, \ \forall k \in \mathbb{N}; \ h_{N_\varepsilon} = 0 \right\}.$$

The class $\mathcal{H}_{\mathrm{mon}}$ is of high interest in statistics because it contains the most common filters such as the projection filter, the Pinsker filter, the Tikhonov or smoothing spline filter and so forth.

Proposition 1. *Set $\gamma_\varepsilon = \max_{1\le j\le J-1}(\sigma_{j+1}^2/\sigma_j^2)$. Under the conditions of Theorem 5, it holds*

$$\varepsilon^{-2}\|f'\|^2\mathbf{E}_{\vartheta,f}[(\hat\vartheta_\varepsilon^S-\vartheta)^2]\le 1+\gamma_\varepsilon(1+\alpha_\varepsilon)\frac{\min_{h\in\mathcal{H}_{\mathrm{mon}}}R^\varepsilon[f,h]}{\|f'\|^2},$$

where $\alpha_\varepsilon\to 0$ as $\varepsilon\to 0$ uniformly in $f\in\mathcal{F}(\beta_*,L_*,\rho)$.

Remark 4. *For the blocks defined by (2.15), we have $T_\varepsilon=\nu_\varepsilon\rho_\varepsilon(1+\rho_\varepsilon)$, $\sigma_1^2\le 4\pi^2\nu_\varepsilon^3$ and $-\nu_\varepsilon\rho_\varepsilon+\nu_\varepsilon(1+\rho_\varepsilon)^j\le\kappa_{j+1}\le 1+\nu_\varepsilon(1+\rho_\varepsilon)^j$. One also checks that $\gamma_\varepsilon=\max_j\sigma_{j+1}^2/\sigma_j^2$ is asymptotically equivalent to $(1+\rho_\varepsilon)^3\sim 1+3\rho_\varepsilon$ as $\varepsilon\to 0$. Therefore the factor in the oracle inequality of Proposition 1 is of order $(1+3\rho_\varepsilon+\alpha_\varepsilon)$. We have already mentioned that $\alpha_\varepsilon^2=O(T_\varepsilon^{-1}\log\varepsilon^{-1})$. The trade-off between α_ε and ρ_ε leads us to $\rho_\varepsilon\asymp\nu_\varepsilon^{-1/3}$. This clarifies our choice of ρ_ε, which is slightly different from the one of [14].*

Remark 5. *In [13, 44, 51], the weakly geometrically increasing blocks are defined by $T_j=\lfloor\nu(1+\rho)^{j-1}\rfloor$. This type of blocks does not lead to a sharp oracle inequality in our case, since we need not only $\max(T_{j+1}/T_j)\to 1$, but also $\max(\kappa_{j+1}/\kappa_j)\to 1$ as $\varepsilon\to 0$.*

To complete the theoretical analysis, we state the result assessing that the estimator $\hat\vartheta_\varepsilon^S$ corresponding to the blocks (2.15) enjoys minimax properties over a large scale of Sobolev balls. Assume that $\bar f\in\mathcal{F}(\beta^*,L^*,\rho)$ and define

$$\mathcal{F}_{\delta,\beta,L}(\bar f)=\Big\{f=\bar f+v:\ \|v\|\le\delta,\ \|v^{(\beta)}\|\le L\Big\}.$$

Theorem 6. *Assume that the conditions of Theorem 5 are fulfilled. If $\delta=\delta_\varepsilon$ tends to zero as $\varepsilon\to 0$ and $\bar f\in\mathcal{F}(\beta^*,L^*,\rho)$ with $\beta^*>\beta\ge\beta_*$, then*

$$\sup_{\vartheta\in\Theta,f\in\mathcal{F}_{\delta,\beta,L}(\bar f)}I^\varepsilon(f)\mathbf{E}_{\vartheta,f}[(\hat\vartheta_\varepsilon^S-\vartheta)^2]\le 1+(1+o(1))\frac{\tilde P(\beta,L)\varepsilon^{\frac{4\beta-4}{2\beta+1}}}{\|\bar f'\|^2},$$

when $\varepsilon\to 0$, with $\tilde P(\beta,L)=\frac{1}{3}\big(\frac{\beta-1}{2\pi(\beta+2)}\big)^{\frac{2\beta-2}{2\beta+1}}(L(2\beta+1))^{\frac{3}{2\beta+1}}$. Moreover, the following lower bound holds:

$$\inf_{\tilde\vartheta_\varepsilon}\sup_{\vartheta\in\Theta,f\in\mathcal{F}_{\delta,\beta,L}(\bar f)}I^\varepsilon(f)\mathbf{E}_{\vartheta,f}[(\tilde\vartheta_\varepsilon-\vartheta)^2]\ge 1+(1+o(1))\frac{\tilde P(\beta,L)\varepsilon^{\frac{4\beta-4}{2\beta+1}}}{\|\bar f'\|^2},$$

where the inf *is taken over all possible estimators $\tilde\vartheta_\varepsilon$.*

2.3 Dimension reduction for nonparametric regression

Throughout this section we assume that we are given n observations $(Y_1,X_1),\dots,(Y_n,X_n)\in\mathbb{R}\times\mathbb{R}^d$ from the model

$$Y_i=f(X_i)+\xi_i=g(\vartheta_1^\top X_i,\dots,\vartheta_{m^*}^\top X_i)+\xi_i, \tag{2.17}$$

where $\xi_1, \dots, \xi_n$ are unobserved errors assumed to be mutually independent zero mean random variables, independent of the design $\{X_i, i \le n\}$.

We are interested in the problem of estimating the subspace $\mathcal{S}_\vartheta = \mathrm{Span}(\vartheta_k, k \le m^*)$. In general, for a fixed function f, there are many ways of choosing g, m^* and $\{\vartheta_k, k \le m^*\}$ so that $f(X_i) = g(\vartheta_1^\top X_i, \dots, \vartheta_{m^*}^\top X_i)$ for every $i = 1, \dots, n$. To avoid this unidentifiability issue, we focus our attention on the estimation of the minimal subspace $\mathcal{S}$, which is the intersection of all linear subspaces S_0 such that, for every $i = 1, \dots, n$, the value of f at X_i depends only on the projection of X_i on $\mathcal{S}_0$. One easily checks that $\mathcal{S}$ coincides with the range of the matrix $\boldsymbol{\nabla f} = (\partial_j f(X_i))_{j \le d, i \le n}$. The subspace $\mathcal{S}$ we wish to estimate is called effective dimension-reduction (EDR) subspace. In what follows, we assume that the design $\{X_i,\ i \le n\}$ is frozen and deterministic, so $\mathcal{S}$ is deterministic as well. We use the notation $X_{ij} = X_i - X_j$.

The Structure-Adaptive algorithm with Maximum Minimization (SAMM) consists of following steps.

a) Specify positive real numbers a_ρ, a_h, ρ_1 and h_1. Choose an integer L and select a set $\{\psi_\ell,\ \ell \le L\}$ of vectors from $\mathbb{R}^n$ verifying $|\psi_\ell|^2 = n$. Set $k = 1$.

b) Initialize the parameters $h = h_1$, $\rho = \rho_1$ and $\widehat{\Pi}_0 = 0$.

c) Define the estimators $\widehat{\nabla f}(X_i)$ for $i = 1, \dots, n$ by formula

$$\begin{pmatrix} \hat{f}(X_i) \\ \widehat{\nabla f}(X_i) \end{pmatrix} = \Big\{ \sum_{j=1}^n \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix}^\top w_{ij} + I_{d+1}/n \Big\}^{-1} \sum_{j=1}^n Y_j \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} w_{ij} \ .$$

where $w_{ij} = K(X_{ij}^\top (I + \rho^{-2}\widehat{\Pi}) X_{ij}/h^2)$ with the current values of h, ρ and $\widehat{\Pi}$. Set

$$\hat{\beta}_\ell = \frac{1}{n} \sum_{i=1}^n \widehat{\nabla f}(X_i) \psi_{\ell,i}, \qquad \ell = 1, \dots, L,$$

where $\psi_{\ell,i}$ is the ith coordinate of ψ_ℓ.

d) Define the new value $\widehat{\Pi}_k$ by $\widehat{\Pi}_k \in \arg\min_{\Pi \in \mathcal{A}_{m^*}} \max_\ell \hat{\beta}_\ell^\top (I - \Pi) \hat{\beta}_\ell$, where $\mathcal{A}_{m^*} = \{\Pi : \Pi = \Pi^\top,\ 0 \preceq \Pi \preceq I,\ \mathrm{Tr}(\Pi) \le m^*\}$.

e) Set $\rho_{k+1} = a_\rho \cdot \rho_k$, $h_{k+1} = a_h \cdot h_k$ and increase k by one.

f) Stop if $\rho < \rho_{\min}$ or $h > h_{\max}$, otherwise continue with the step c).

Let $k(n)$ be the total number of iterations. The matrix $\widehat{\Pi}_{k(n)}$ is the desired estimator of the projector Π^*. We denote by $\widehat{\Pi}_n$ the orthogonal projection onto the space spanned by the eigenvectors of $\widehat{\Pi}_{k(n)}$ corresponding to the m^* largest eigenvalues. The estimator of the EDR subspace is then the image of $\widehat{\Pi}_n$. Equivalently, $\widehat{\Pi}_n$ is the estimator of the projector onto $\mathcal{S}$.

The described algorithm requires the specification of the parameters ρ_1, h_1, a_ρ and a_h, as well as the choice of the set of vectors $\{\psi_\ell\}$. In what follows we use the values

$$\begin{gathered} \rho_1 = 1, \qquad \rho_{\min} = n^{-1/(3 \vee m^*)},\ a_\rho = e^{-1/2(3 \vee m^*)}, \\ h_1 = C_0 n^{-1/(4 \vee d)}, \quad h_{\max} = 2\sqrt{d}, \qquad a_h = e^{1/2(4 \vee d)}. \end{gathered}$$

In our assumptions we will implicitly assume that the neighborhood $E^{(k)}(X_i) = \{x : |(I + \rho_k^{-2}\Pi^*)^{-1/2}(X_i - x)| \le h_k\}$ contains at least d design points different from X_i. The parameters h_1, ρ_1, a_ρ and a_h are chosen so that the volume of ellipsoids $E^{(k)}(X_i)$ is a non-decreasing function of k and $Vol(E^{(1)}(X_i)) = C_0/n$. In applications, we define h_1 as the smallest real such that $\min_{i=1,\dots,n} \#E^{(1)}(X_i) = d + 1$.

The set $\{\psi_\ell\}$ plays an essential role in the algorithm. The optimal choice of this set is an important issue that needs further investigation. We content ourselves with giving one particular choice which agrees with theory and leads to nice empirical results. Let $\mathfrak{S}_j$, $j \le d$, be the permutation of the set $\{1, \dots, n\}$ satisfying $X^{(j)}_{\mathfrak{S}_j(1)} \le \dots \le X^{(j)}_{\mathfrak{S}_j(n)}$. Let $\mathfrak{S}_j^{-1}$ be the inverse of $\mathfrak{S}_j$, i.e. $\mathfrak{S}_j(\mathfrak{S}_j^{-1}(k)) = k$ for every $k = 1, \dots, n$. Define $\{\psi_\ell\}$ as the set of vectors

$$\left\{ \begin{array}{c} \Big(\cos\big(\frac{2\pi(k-1)\mathfrak{S}_j^{-1}(1)}{n}\big), \dots, \cos\big(\frac{2\pi(k-1)\mathfrak{S}_j^{-1}(n)}{n}\big) \Big)^\top \\ \Big(\sin\big(\frac{2\pi k\mathfrak{S}_j^{-1}(1)}{n}\big), \dots, \sin\big(\frac{2\pi k\mathfrak{S}_j^{-1}(n)}{n}\big) \Big)^\top \end{array} , k \le [n/2],\ j \le d \right\}$$

normalized to satisfy $\sum_{i=1}^n \psi_{\ell,i}^2 = n$ for every ℓ. Above, $[n/2]$ is the integer part of $n/2$ and k and j are positive integers.

Theorem 7. *Assume that assumptions [P9, **(A1)-(A4)**] are fulfilled. There exists a constant $C > 0$ such that for any $z \in]0, 2\sqrt{\log(nL)}]$ and for sufficiently large values of n, it holds*

$$\mathbf{P}\left(\|\widehat{\Pi}_n - \Pi^*\|_2 > Cn^{-\frac{2}{3\vee m^*}} t_n^2 + \frac{Cz\sqrt{d}}{\sqrt{n(1-\zeta_n)}} \right) \le Lze^{-\frac{z^2-1}{2}} + \frac{3k(n)-5}{n},$$

where $t_n = O(\sqrt{\log(Ln)})$ and $\zeta_n = O(t_n\, n^{-\frac{1}{6\vee m^*}})$.

This result assesses that for $m^* \le 4$, the estimator of $\mathcal{S}$ provided by the SAMM procedure is $\sqrt{n}$-consistent up to a logarithmic factor. This rate of convergence is known to be optimal for a broad class of semiparametric problems.

Let us present now the results of some simulations. In all examples presented below the number of replications is $N = 250$. The mean loss $\overline{\mathrm{er}_N} = \frac{1}{N}\sum_j \mathrm{er}_j$ and the standard deviation $\sqrt{\frac{1}{N}\sum_j (\mathrm{er}_j - \overline{\mathrm{er}_N})^2}$ are reported, where $\mathrm{er}_j = \|\widehat{\Pi}^{(j)} - \Pi^*\|$ with $\widehat{\Pi}^{(j)}$ being the estimator of Π^* for jth replication.

Example 1 (Single-index). We set $d = 5$ and $f(x) = g(\vartheta^\top x)$ with

$$g(t) = 4|t|^{1/2}\sin^2(\pi t), \quad \text{and} \quad \vartheta = (1/\sqrt{5}, 2/\sqrt{5}, 0, 0, 0)^\top \in \mathbb{R}^5.$$

We run SAMM and MAVE procedures on the data generated by the model

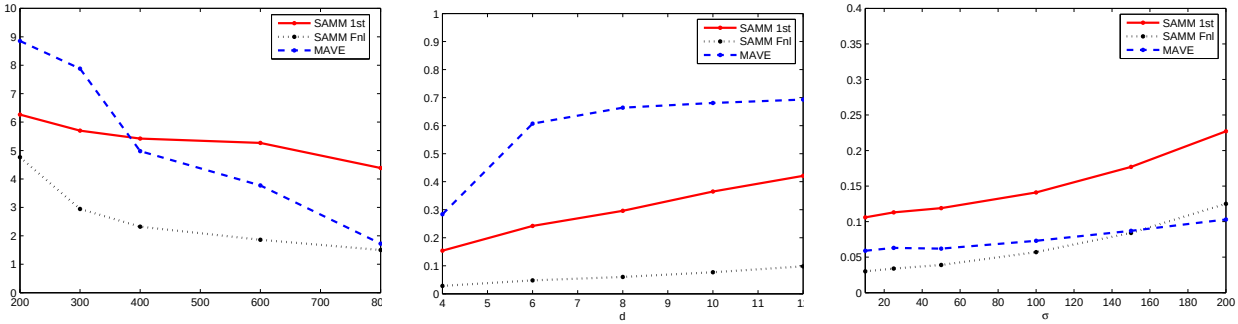
$$Y_i = f(X_i) + 0.5 \cdot \xi_i,$$

where the design X is such that the coordinates $(X_i^{(j)}, j \le 5, i \le n)$ are i.i.d. uniform on $[-1, 1]$, and the errors ξ_i are i.i.d. standard Gaussian independent of the design.

Table 2.1 contains the average loss for different values of the sample size n for the first step estimator by SAMM, the final estimator provided by SAMM and the estimator based on

Table 2.1: *Average loss* $\|\widehat{\Pi} - \Pi^*\|$ *of the estimators obtained by SAMM and MAVE procedures in Example 1. The standard deviation is given in parentheses.*

n	200	300	400	600	800
SAMM, 1st	0.443	0.329	0.271	0.215	0.155
	(.211)	(.120)	(.115)	(.095)	(.079)
SAMM, Fnl	0.337	0.170	0.116	0.076	0.053
	(.273)	(.147)	(.104)	(.054)	(.031)
MAVE	0.626	0.455	0.249	0.154	0.061
	(.363)	(.408)	(.342)	(.290)	(.161)



(a) **(b)** **(c)**

Figure 2.1: (a) Average loss multiplied by $\sqrt{n}$ versus n for the first step (full line) and the final (dotted line) estimators provided by SAMM and for the estimator based on MAVE (broken line) in Example 1, (b) (resp. (c)) Average loss versus d (resp. σ) for the first step (full line) and the final (dotted line) estimators provided by SAMM and for the estimator based on MAVE (broken line) in Example 2 (resp. Example 3).

MAVE. We plot in Figure 2.1 (a) the average loss normalized by the square rood of the sample size n versus n. It is clearly seen that the iterative procedure improves considerably the quality of estimation and that the final estimator provided by SAMM is $\sqrt{n}$-consistent. In this example, MAVE method often fails to recover the EDR subspace. However, the number of failures decreases very rapidly with increasing n. This is the reason why the curve corresponding to MAVE in Figure 2.1 (a) decreases with a strong slope.

Example 2 (Double-index). For $d \geq 2$ we set $f(x) = g(\vartheta^\top x)$ with

$$g(x) = (x_1 - x_2^3)(x_1^3 + x_2);$$

and $\vartheta_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$, $\vartheta_2 = (0, 1, \ldots, 0) \in \mathbb{R}^d$. We run SAMM and MAVE procedures on the data generated by the model

$$Y_i = f(X_i) + 0.1 \cdot \xi_i, \qquad i = 1, \ldots, 300,$$

where the design X is such that the coordinates $(X_i^{(j)}, j \leq d, i \leq n)$ are i.i.d. uniform on $[-40, 40]$, and the errors ξ_i are i.i.d. standard Gaussian independent of the design. The results of simulations for different values of d are reported in Table 2.2.

As expected, we found that (cf. Figure 2.1(b)) the quality of SAMM deteriorated linearly in d as d increased. This agrees with our theoretical results. It should be noted that in this case MAVE fails to find the EDR space.

Table 2.2: *Average loss* $\|\widehat{\Pi} - \Pi^*\|$ *of the estimators obtained by SAMM and MAVE procedures in Example 2. The standard deviation is given in parentheses.*

d	4	6	8	10	12
SAMM 1st	0.154	0.242	0.296	0.365	0.421
	(.063)	(.081)	(.071)	(.087)	(.095)
SAMM, Fnl	0.028	0.048	0.060	0.077	0.098
	(.011)	(.020)	(.021)	(.026)	(.037)
MAVE	0.284	0.607	0.664	0.681	0.693
	(.147)	(.073)	(.052)	(.054)	(.044)

Table 2.3: *Average loss* $\|\widehat{\Pi} - \Pi^*\|$ *of the estimators obtained by SAMM and MAVE procedures in Example 3. The standard deviation is given in parentheses.*

σ	200	150	100	50	25	10
SAMM 1st	0.227	0.177	0.141	0.119	0.113	0.106
	(.092)	(.075)	(.055)	(.051)	(.048)	(.043)
SAMM, Fnl	0.125	0.084	0.057	0.039	0.034	0.030
	(.076)	(.037)	(.026)	(.019)	(.021)	(.018)
MAVE	0.103	0.087	0.073	0.062	0.063	0.059
	(.041)	(.035)	(.027)	(.023)	(.024)	(.023)

Example 3. For $d = 5$ we set $f(x) = g(\vartheta^\top x)$ with

$$g(x) = (1 + x_1)(1 + x_2)(1 + x_3)$$

and $\vartheta_1 = (1, 0, 0, 0, 0)$, $\vartheta_2 = (0, 1, 0, 0, 0)$, $\vartheta_3 = (0, 0, 1, 0, 0)$. We run SAMM and MAVE procedures on the data generated by the model

$$Y_i = f(X_i) + \sigma \cdot \xi_i, \qquad i = 1, \dots, 250,$$

where the design X is such that the coordinates $(X_i^{(j)}, j \le d, i \le n)$ are i.i.d. uniform on $[0, 20]$, and the errors ξ_i are i.i.d. standard Gaussian independent of the design.

Figure 2.1(c) shows that the qualities of both SAMM and MAVE deteriorate linearly in σ, when σ increases. These results also demonstrate that, thanks to an efficient bias reduction, the SAMM procedure outperforms MAVE when stochastic error is small, whereas MAVE works better than SAMM in the case of dominating stochastic error (that is when σ is large).

2.4 Aggregation for nonparametric regression

Consider the regression model

$$Y_i = f(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{2.18}$$

where $x_1, \dots, x_n$ are given elements of a set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$ is an unknown function, and ξ_i are i.i.d. zero-mean random variables on a probability space $(\Omega, \mathcal{F}, P)$ where $\Omega \subseteq \mathbb{R}$. The problem is to estimate the function f from the data $D_n = ((x_1, Y_1), \dots, (x_n, Y_n))$.

Let $(\Lambda, \mathcal{A})$ be a probability space and denote by $\mathscr{P}_\Lambda$ the set of all probability measures defined on $(\Lambda, \mathcal{A})$. Assume that we are given a family $\{f_\lambda, \lambda \in \Lambda\}$ of functions $f_\lambda : \mathcal{X} \to \mathbb{R}$ such that the mapping $\lambda \mapsto f_\lambda$ is measurable, $\mathbb{R}$ being equipped with the Borel σ-field. Functions f_λ can be viewed either as weak learners or as some preliminary estimators of f based on a training sample independent of $\mathbf{Y} \triangleq (Y_1, \dots, Y_n)$ and considered as frozen.

We study the problem of aggregation of functions in $\{f_\lambda, \lambda \in \Lambda\}$ under the squared loss. Specifically, we construct an estimator $\hat{f}_n$ based on the data D_n and called *aggregate* such that the expected value of its squared error

$$\|\hat{f}_n - f\|_n^2 \triangleq \frac{1}{n} \sum_{i=1}^n \big(\hat{f}_n(x_i) - f(x_i)\big)^2$$

is approximately as small as the oracle value $\inf_{\lambda \in \Lambda} \|f - f_\lambda\|_n^2$.

In [P10], we consider aggregates that are mixtures of functions f_λ with exponential weights. For a measure π from $\mathscr{P}_\Lambda$ and for $\beta > 0$ we set

$$\hat{f}_n(x) \triangleq \int_\Lambda \vartheta_\lambda(\beta, \pi, \mathbf{Y}) f_\lambda(x)\, \pi(d\lambda), \quad x \in \mathcal{X}, \tag{2.19}$$

with

$$\vartheta_\lambda(\beta, \pi, \mathbf{Y}) = \frac{\exp\big\{-n\|\mathbf{Y} - f_\lambda\|_n^2/\beta\big\}}{\int_\Lambda \exp\big\{-n\|\mathbf{Y} - f_w\|_n^2/\beta\big\}\pi(dw)} \tag{2.20}$$

where $\|\mathbf{Y} - f_\lambda\|_n^2 \triangleq \frac{1}{n}\sum_{i=1}^n \big(Y_i - f_\lambda(x_i)\big)^2$ and we assume that π is such that the integral in (2.19) is finite.

Note that $\hat{f}_n$ depends on two tuning parameters: the prior measure π and the "temperature" parameter β. They have to be selected in a suitable way. Using the Bayesian terminology, $\pi(\cdot)$ is a prior distribution and $\hat{f}_n$ is the posterior mean of f_λ in a "phantom" model $Y_i = f_\lambda(x_i) + \xi_i'$, where ξ_i' are iid normally distributed with mean 0 and variance $\beta/2$.

Our assumptions concern essentially the probability distribution of the i.i.d. errors ξ_i.

(A) There exist i.i.d. random variables $\zeta_1, \dots, \zeta_n$ defined on an enlargement of the probability space $(\Omega, \mathcal{F}, P)$ such that:

(A1) the random variable $\xi_1 + \zeta_1$ has the same distribution as $(1 + 1/n)\xi_1$,

(A2) the vectors $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ are independent.

Note that (A) is an assumption on the distribution of ξ_1. If ξ_1 satisfies (A1), then we will say that its distribution is *n-divisible*.

Hereafter, we will write for brevity ϑ_λ instead of $\vartheta_\lambda(\beta, \pi, \mathbf{Y})$. Denote by $\mathscr{P}'_\Lambda$ the set of all the measures $\mu \in \mathscr{P}_\Lambda$ such that $\lambda \mapsto f_\lambda(x)$ is integrable w.r.t. μ for $x \in \{x_1, \dots, x_n\}$. Clearly $\mathscr{P}'_\Lambda$

is a convex subset of $\mathscr{P}_\Lambda$. For any measure $\mu \in \mathscr{P}'_\Lambda$ we define

$$\bar{f}_\mu(x_i) = \int_\Lambda f_\lambda(x_i)\, \mu(d\lambda), \quad i = 1, \dots, n.$$

We denote by $\vartheta \cdot \pi$ the probability measure $A \mapsto \int_A \vartheta_\lambda \, \pi(d\lambda)$ defined on $\mathcal{A}$. With the above notation, we have $\hat{f}_n = \bar{f}_{\vartheta \cdot \pi}$.

We will need one more assumption. Let $L_\zeta : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be the moment generating function of the random variable ζ_1, i.e., $L_\zeta(t) = E(e^{t\zeta_1})$, $t \in \mathbb{R}$.

(B) There exist a functional $\Psi_\beta : \mathscr{P}'_\Lambda \times \mathscr{P}'_\Lambda \to \mathbb{R}$ and a real number $\beta_0 > 0$ such that

$$\begin{cases} e^{(\|f-\bar{f}_{\mu'}\|_n^2 - \|f-\bar{f}_\mu\|_n^2)/\beta} \prod_{i=1}^n L_\zeta\Big(\frac{2(\bar{f}_\mu(x_i) - \bar{f}_{\mu'}(x_i))}{\beta}\Big) \le \Psi_\beta(\mu, \mu'), \\ \mu \mapsto \Psi_\beta(\mu, \mu') \text{ is concave and continuous in the total} \\ \text{variation norm for any } \mu' \in \mathscr{P}'_\Lambda, \\ \Psi_\beta(\mu, \mu) = 1, \end{cases} \tag{2.21}$$

for any $\beta \ge \beta_0$.

Theorem 8. *Let π be an element of $\mathscr{P}_\Lambda$ such that $\vartheta \cdot \pi \in \mathscr{P}'_\Lambda$ for all $\mathbf{Y} \in \mathbb{R}^n$ and $\beta > 0$. If assumptions (A) and (B) are fulfilled, then the aggregate $\hat{f}_n$ defined by (2.19) with $\beta \ge \beta_0$ satisfies the oracle inequality*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \int \|f_\lambda - f\|_n^2 \, p(d\lambda) + \frac{\beta\, \mathcal{K}(p, \pi)}{n+1}, \quad \forall\, p \in \mathscr{P}_\Lambda, \tag{2.22}$$

where $\mathcal{K}(p, \pi)$ stands for the Kullback-Leibler divergence between p and π.

Consider now the particular case where Λ is countable. W.l.o.g. we suppose that $\Lambda = \{1, 2, \dots\}$, $\{f_\lambda, \lambda \in \Lambda\} = \{f_j\}_{j=1}^\infty$ and we set $\pi_j \triangleq \pi(\lambda = j)$. As a corollary of Theorem 8 we get the following sharp oracle inequalities for model selection type aggregation.

Theorem 9. *Assume that π is an element of $\mathscr{P}_\Lambda$ such that $\vartheta \cdot \pi \in \mathscr{P}'_\Lambda$ for all $\mathbf{Y} \in \mathbb{R}^n$ and $\beta > 0$. Let assumptions (A) and (B) be fulfilled and let Λ be countable. Then for any $\beta \ge \beta_0$ the aggregate $\hat{f}_n$ satisfies the inequality*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \inf_{j \ge 1} \left(\|f_j - f\|_n^2 + \frac{\beta \log \pi_j^{-1}}{n+1} \right).$$

In particular, if $\pi_j = 1/M$, $j = 1, \dots, M$, we have

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \min_{j=1,\dots,M} \|f_j - f\|_n^2 + \frac{\beta \log M}{n+1}. \tag{2.23}$$

The rate of convergence $(\log M)/n$ obtained in (2.23) is optimal rate of model selection type aggregation when the errors ξ_i are Gaussian.

We now discuss two important cases of Theorem 8 where the errors ξ_i are either Gaussian or double exponential.

Proposition 2. *Assume that* $\sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n \le L < \infty$. *If the random variables* ξ_i *are i.i.d. Gaussian* $\mathcal{N}(0,\sigma^2)$, $\sigma^2 > 0$, *then for every* $\beta \ge (4+2/n)\sigma^2 + 2L^2$ *the aggregate* $\hat{f}_n$ *satisfies inequality (2.22).*

Assume now that ξ_i are distributed with the double exponential density

$$f_\xi(x) = \frac{1}{\sqrt{2\sigma^2}}\, e^{-\sqrt{2}|x|/\sigma}, \quad x \in \mathbb{R}.$$

Aggregation under this assumption is discussed in [56] where it is recommended to modify the shape of the aggregate in order to match the shape of the distribution of the errors. The next proposition shows that sharp risk bounds can be obtained without modifying the algorithm.

Proposition 3. *Assume that* $\sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n \le L < \infty$ *and* $\sup_{i,\lambda}|f_\lambda(x_i)| \le \bar{L} < \infty$. *Let the random variables* ξ_i *be i.i.d. double exponential with variance* $\sigma^2 > 0$. *Then for any* β *larger than*

$$\max\left(\left(8+\frac{4}{n}\right)\sigma^2 + 2L^2,\ 4\sigma\left(1+\frac{1}{n}\right)\bar{L}\right)$$

the aggregate $\hat{f}_n$ *satisfies inequality (2.22).*

As discussed above, assumption (A) restricts the application of Theorem 8 to models with "n-divisible" errors. Using a construction inspired by the Skorokhod embedding, we succeed in extending the desired oracle inequality to a wider class of noise distributions. For simplicity we assume that the errors ξ_i are symmetric, i.e., $P(\xi_i > a) = P(\xi_i < -a)$ for all $a \in \mathbb{R}$. The argument can be adapted to the asymmetric case as well, but we do not discuss it here.

Theorem 10. *Fix some* $\alpha > 0$ *and assume that* $\sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n \le L$ *for a finite constant* L. *If the errors* ξ_i *are symmetric and have a finite second moment* $E(\xi_i^2)$, *then for any* $\beta \ge 4(1+1/n)\alpha + 2L^2$ *we have*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \int_\Lambda \|f_\lambda - f\|_n^2\, p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1} + R_n, \quad \forall\, p \in \mathscr{P}_\Lambda, \tag{2.24}$$

where the residual term R_n *is given by*

$$R_n = E^*\left(\sup_{\lambda\in\Lambda}\sum_{i=1}^n \frac{4(n+1)(\xi_i^2-\alpha)(f_\lambda(x_i)-\bar{f}_{\vartheta\cdot\pi}(x_i))^2}{n^2\beta}\right)$$

and E^* *denotes expectation with respect to the outer probability* P^*.

Corollary 1. *Let the assumptions of Theorem 10 be satisfied and let* $|\xi_i| \le B$ *almost surely where* B *is a finite constant. Then the aggregate* $\hat{f}_n$ *satisfies inequality (2.22) for any* $\beta \ge 4B^2(1+1/n) + 2L^2$.

Corollary 2. *Let the assumptions of Theorem 10 be satisfied and suppose that* $E(e^{t|\xi_i|^\kappa}) \le B$ *for some finite constants* $t > 0$, $\kappa > 0$, $B > 0$. *Then for any* $n \ge e^{2/\kappa}$ *and any* $\beta \ge 4(1+1/n)(2(\log n)/t)^{1/\kappa} + 2L^2$ *we have*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \int_\Lambda \|f_\lambda - f\|_n^2\, p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1} + \frac{16BL^2(n+1)(2\log n)^{2/\kappa}}{n^2\beta\, t^{2/\kappa}}, \quad \forall\, p \in \mathscr{P}_\Lambda.$$

In particular, if $\Lambda = \{1,\dots,M\}$ *and* π *is the uniform measure on* Λ *we get*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \min_{j=1,\dots,M}\|f_j - f\|_n^2 + \frac{\beta\,\log M}{n+1} + \frac{16BL^2(n+1)(2\log n)^{2/\kappa}}{n^2\beta\, t^{2/\kappa}}.$$

Interestingly, the obtained results can be used to derive sparsity oracle inequalities. Let $\phi_1, \dots, \phi_M$ be some functions from $\mathcal{X}$ to $\mathbb{R}$. Consider the case where $\Lambda \subseteq \mathbb{R}^M$ and $f_\lambda = \sum_j \lambda_j \phi_j$, $\lambda = (\lambda_1, \dots, \lambda_M)$. For $\lambda \in \mathbb{R}^M$ denote by $J(\lambda)$ the set of indices j such that $\lambda_j \neq 0$, and set $M(\lambda) \triangleq Card(J(\lambda))$. For any $\tau > 0$, $0 < L_0 \leq \infty$, define the probability densities

$$q_0(t) = \frac{3}{2(1+|t|)^4}, \ \forall t \in \mathbb{R},$$
$$q(\lambda) = \frac{1}{C_0} \prod_{j=1}^{M} \tau^{-1} \, q_0\big(\lambda_j/\tau\big) \mathbf{1}(\|\lambda\| \leq L_0), \ \forall \lambda \in \mathbb{R}^M,$$

where $C_0 = C_0(\tau, M, L_0)$ is the normalizing constant and $\|\lambda\|$ stands for the Euclidean norm of $\lambda \in \mathbb{R}^M$.

Sparsity oracle inequalities (SOI) are oracle inequalities bounding the risk in terms of the sparsity index $M(\lambda)$ or similar characteristics. The next theorem provides a general tool to derive SOI from the "PAC-Bayesian" bound (2.22). Note that in this theorem $\hat{f}_n$ is not necessarily defined by (2.19). It can be any procedure satisfying (2.22).

Theorem 11. *Let $\hat{f}_n$ satisfy (2.22) with $\pi(d\lambda) = q(\lambda)\, d\lambda$ and $\tau \leq \delta L_0/\sqrt{M}$ where $0 < L_0 \leq \infty$, $0 < \delta < 1$. Assume that Λ contains the ball $\{\lambda \in \mathbb{R}^M : \|\lambda\| \leq L_0\}$. Then for all λ^* such that $\|\lambda^*\| \leq (1-\delta)L_0$ we have*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \leq \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta}{n+1} \sum_{j \in J(\lambda^*)} \log(1 + \tau^{-1}|\lambda_j^*|) + R(M, \tau, L_0, \delta),$$

where the residual term is

$$R(M, \tau, L_0, \delta) = \tau^2 e^{2\tau^3 M^{5/2} (\delta L_0)^{-3}} \sum_{j=1}^{M} \|\phi_j\|_n^2 + \frac{2\beta\tau^3 M^{5/2}}{(n+1)\delta^3 L_0^3}$$

for $L_0 < \infty$ and $R(M, \tau, \infty, \delta) = \tau^2 \sum_{j=1}^{M} \|\phi_j\|_n^2$.

We now discuss a consequence of the obtained inequality in the case where the errors are Gaussian. Let us denote by Φ the Gram matrix associated to the family $(\phi_j)_{j=1,\dots,M}$, i.e., $M \times M$ matrix with entries $\Phi_{j,j'} = n^{-1} \sum_{i=1}^{n} \phi_j(x_i)\phi_{j'}(x_i)$ for every $j, j' \in \{1, \dots, M\}$. We denote by $\lambda_{\max}(\Phi)$ the maximal eigenvalue of Φ. In what follows, for every $x > 0$, we write $\log_+ x = (\log x)_+$.

Corollary 3. *Let $\hat{f}_n$ be defined by (2.19) with $\pi(d\lambda) = q(\lambda)\, d\lambda$ and let $\tau = \frac{\delta L_0}{M\sqrt{n}}$ with $0 < L_0 < \infty$, $0 < \delta < 1$. Let ξ_i be i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 > 0$, $\lambda_{max}(\Phi) \leq K^2$, $\|f\|_n \leq \bar{L}$ and let $\beta \geq (4 + 2n^{-1})\sigma^2 + 2L^2$ with $L = \bar{L} + L_0 K$. Then for all $\lambda^* \in \mathbb{R}^M$ such that $\|\lambda^*\| \leq (1-\delta)L_0$ we have*

$$E\big[\|\hat{f}_n - f\|_n^2\big] \leq \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta}{n+1}\Big[M(\lambda^*)\Big(1 + \log_+\Big\{\frac{M\sqrt{n}}{\delta L_0}\Big\}\Big) + \sum_{J(\lambda^*)} \log_+ |\lambda_j^*|\Big]$$
$$+ \frac{C}{nM^{1/2}\min(M^{1/2}, n^{3/2})},$$

where C is a positive constant independent of n, M and λ^*.

Publications

[P1] DALALYAN, A. S. AND KUTOYANTS, YU. A.: Asymptotically efficient estimation of the derivative of the invariant density. *Stat. Inference Stoch. Process.* **6** (2003), no. 1, 89–107.

[P2] DALALYAN, A. S. AND KUTOYANTS, YU. A.: Asymptotically efficient trend coefficient estimation for ergodic diffusion. *Math. Methods of Statist.* **11** (2002), no. 4, 402–427.

[P3] DALALYAN, A. S. AND KUTOYANTS, YU. A.: On second order minimax estimation of invariant density for ergodic diffusion. STATIST. DECISIONS **22** (2004), no. 1, 17–41.

[P4] DALALYAN, A.: Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.* **33** (2005), no. 6, 2507–2528.

[P5] DALALYAN, A. AND REISS, M.: Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields* **134** (2006), no. 2, 248–282.

[P6] DALALYAN, A. AND REISS, M.: Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case. *Probab. Theory Related Fields* **137** (2007), no. 1-2, 25–47.

[P7] DALALYAN, A. S.; GOLUBEV, G. K. AND TSYBAKOV, A. B.: Penalized maximum likelihood and semiparametric second-order efficiency. *Ann. Statist.* **34** (2006), no. 1, 169–201.

[P8] DALALYAN, A. S.: Stein shrinkage and second-order efficiency for semiparametric estimation of the shift. *Math. Methods of Statist.* **16** (2007), no. 1, 43–63.

[P9] DALALYAN, A. S.; JUDITSKY, A. AND SPOKOINY, V.: A new algorithm for estimating the effective dimension-reduction subspace, *submitted*.

[P10] DALALYAN, A. S. AND TSYBAKOV, A. B.: Aggregation by exponential weighting and sharp oracle inequalities. *Proceedings of The Twentieth Annual Conference on Learning Theory (COLT 2007).* (This paper is invited by the *Machine Learning* journal.)

Bibliography

[1] AÏT-SAHALIA, Y.: Transition Densities for Interest Rate and Other Nonlinear Diffusions. *Journal of Finance*, **54** (1999), 1361–1395.

[2] AÏT-SAHALIA, Y.: Closed-form likelihood expansions for multivariate diffusions. Preprint (2004), available under http://www.princeton.edu/∼yacine/research.htm.

[3] AÏT-SAHALIA, Y. AND MYKLAND, P.: Estimating Diffusions with Discretely and Possibly Randomly Spaced Data: A General Theory. *Ann. Statist.* **32** (2004), 2186–2222.

[4] BANDI, F. M., MOLOCHE, G.: On the functional estimation of multivariate diffusion processes. Available under http://gsbwww.uchicago.edu/fac/federico.bandi/research.

[5] BANON, G. Nonparametric identification for diffusion processes. *SIAM J. Control and Optim.* **16** (1978), no. 3, 380–395.

[6] BHATTACHARYA, R. N.: Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Probab.* **6** (1978), no. 4, 541–553.

[7] BLANKE, D. AND BOSQ, D.: Local superefficiency of data-driven projection density estimators in continuous time. *Statist. Oper. Research Trans.* **28** (2004), no. 1, 37–54.

[8] BROWN, LAWRENCE D. AND LOW, MARK G.: Asymptotic equivalence of nonparametric regression and white noise, *Ann. Statist.* **24** (1996), no. 6, 2384–2398.

[9] BROWN, L. D., ZHANG, C.: Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *Ann. Statist.* **26** (1998), 279–287

[10] CARTER, A.: A continuous Gaussian process approximation to a nonparametric regression in two dimensions. BERNOULLI, **12** (2006), no. 1, 143–156.

[11] CASTILLO, I.: Semiparametric second order efficient estimation of the period of a signal. To appear in *Bernoulli*.

[12] CAVALIER, L., GOLUBEV, G. K., PICARD, D. AND TSYBAKOV, A. B.: Oracle inequalities for inverse problems. *Ann. Statist.* **30** (2002), no. 3, 843–874.

[13] CAVALIER, L. AND TSYBAKOV, A.: Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.* **10** (2001), 247–282.

[14] CAVALIER, L. AND TSYBAKOV, A.: Sharp adaptation for inverse problems with random noise. *Proba. Theory and Related Fields* **123** (2002), 323–354.

[15] COOK, R. D.: *Regression graphics. Ideas for studying regressions through graphics.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1998.

[16] COOK, R. D. AND LI, B.: Dimension reduction for conditional mean in regression. *Ann. Statist.* **30** (2002), no. 2, 455–474.

[17] COOK, R. D. AND WEISBERG, S.: Discussion of "Sliced inverse regression for dimension reduction".by K. C. Li, *J. Amer. Statist. Assoc.* **86** (1991), no. 414, 328–332.

[18] COOK, R. D. AND WEISBERG, S.: *Applied Regression Including Computing and Graphics.* Hoboken NJ: John Wiley, 1999.

[19] DELATTRE, S. AND HOFFMANN, M.: Asymptotic equivalence for a null recurrent diffusion. *Bernoulli* **8** (2002), no. 2, 139–174.

[20] FAN, J.: A selective overview of nonparametric methods in financial econometrics. *Statistical Science* **20** (2005), 317–357.

[21] FAN, J. AND GIJBELS, I.: *Local polynomial modelling and its applications.* Monographs on Statistics and Applied Probability, 66. Chapman & Hall, London, 1996.

[22] FAN, J. AND ZHANG, C.: A re-examination of Stanton's diffusion estimations with applications to financial model validation. *J. Amer. Statist. Assoc.* **98** (2003), 118–134.

[23] GIHMAN, I. I. AND SKOROHOD, A. V.: *Stochastic differential equations*. Springer-Verlag, New York-Heidelberg, 1972.

[24] GOLUBEV, G. K.: Non-parametric estimation of smooth densities in L^2. *Problems Inform. Transmission* **28** (1992), 44–54.

[25] GOLUBEV, G. AND HÄRDLE W.: On the second order minimax estimation in partial linear models. *Math. Methods Statist.* **2** (2000), 160–175.

[26] GOLUBEV, G. AND HÄRDLE, W.: On adaptive smoothing in partial linear models. *Math. Methods Statist.* **1** (2002), 98–117.

[27] HOFFMANN, M.: Minimax estimation of the diffusion coefficient through irregular samplings. *Statist. Probab. Lett.* **32** (1997), no. 1, 11–24.

[28] HOFFMANN, M.: Adaptive estimation in diffusion processes. *Stoch. Proc. Appl.* **79** (1999), no. 1, 135–163.

[29] HRISTACHE, M., JUDITSKY, A., POLZEHL, J. AND SPOKOINY, V.: Structure adaptive approach for dimension reduction. *Ann. Statist.* **29** (2001), no. 6, 1537–1566.

[30] JUDITSKY, A., RIGOLLET, P. AND TSYBAKOV, A.: Learning by mirror averaging. Preprint n.1034, Laboratoire de Probabilités et Modèle aléatoires, Universités Paris 6 and Paris 7, 2005. https://hal.ccsd.cnrs.fr/ccsd-00014097

[31] KUTOYANTS, YU. A.: Efficient Density Estimation for Ergodic Diffusion, *Stat. Inference Stoch. Process.* **1** (1998), 131–155.

[32] KUTOYANTS, YU. A.: *Statistical Inference for Ergodic Diffusion Processes,* Springer Series in Statistics, New York, 2003.

[33] LE CAM, L. AND YANG, G. L. *Asymptotics in statistics. Some basic concepts*. Second edition. Springer Series in Statistics. Springer-Verlag, New York, 2000.

[34] LEUNG, G. AND BARRON, A.: Information theory and mixing least-square regressions. *IEEE Transactions on Information Theory* **52** (2006), 3396–3410.

[35] LI, K. C.: Sliced inverse regression for dimension reduction. With discussion and a rejoinder by the author. *J. Amer. Statist. Assoc.* **86** (1991), no. 414, 316–342.

[36] LI, K. C.: On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Amer. Statist. Assoc.*, **87** (1992), 1025–1039.

[37] MILSTEIN, G. AND NUSSBAUM, M.: Diffusion approximation for nonparametric autoregression. *Probab. Theory Related Fields*, **112** (1998), no. 4, 535–543.

[38] MURPHY, S. AND VAN DER VAART, A.: On Profile Likelihood. *J. Amer. Statist. Assoc.* **95** (2000), 449–485.

[39] NUSSBAUM, M.: Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, **24** (1996), no. 6, 2399–2430.

[40] PHAM, T. D.: Nonparametric estimation of the drift coefficient in the diffusion equation, *Math. Operationsforsch. Statist.*, Ser. Statistics, **1** (1981), 61–73.

[41] PINSKER, M. S.: Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16** (1980), 52–68.

[42] REISS, M.: Asymptotic equivalence for nonparametric regression with multivariate and random design, *submitted*.

[43] REVUZ, D. AND YOR, M.: *Continuous Martingales and Brownian Motion.* Third edition. Berlin: Springer-Verlag, 1999.

[44] RIGOLLET, PH.: Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12** (2006), 351–370.

[45] ROGERS, L. C. G., WILLIAMS, D.: *Diffusions, Markov processes, and martingales*. Vol. 2. Itô calculus. Wiley Series in Probability and Mathematical Statistics. New York, 1987.

[46] SEVERINI, T. AND WONG, W.: Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** (1992), 1768–1802.

[47] SPOKOINY, V. G.: Adaptive drift estimation for nonparametric diffusion model, *Ann. Statist.* **28** (2000), 815–836.

[48] STEIN, C.: Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** (1956), 187–196. Univ. of California Press.

[49] STONE, C. J.: Adaptive maximum likelihood estimation of a location parameter. *Ann. Statist.* **3** (1975), 267–284.

[50] TSYBAKOV, A.B.: Regularization, boosting and mirror averaging. Comments on "Regularization in Statistics", by P.Bickel and B.Li. *Test* **15** (2006) 303–310.

[51] TSYBAKOV, A.B.: *Introduction à l'estimation non-paramétrique.* Mathématiques & Applications, 41. Springer-Verlag, Berlin, 2004.

[52] VAN DER VAART, A.: *Semiparametric Statistics,* manuscript downloadable from http://www.math.vu.nl/sto/publications.php

[53] XIA, Y., TONG, H., LI, W. K. AND ZHU, L. X.: An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** (2002), no. 3, 363–410.

[54] YANG, Y.: Combining Different Procedures for Adaptive Regression, *J. Multivariate Anal.* **74** (2000), no. 1, 135–161.

[55] YANG, Y.: Adaptive regression by mixing. *Journal of the American Statistical Association* **96** (2001), 574–588.

[56] YANG, Y.: Regression with multiple candidate models: selecting or mixing? *Statist. Sinica* **13** (2003), 783–809.